Explaining the Behavior of Remote Robots to Humans : An Agent-based Approach

Yazan Mualla

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

utbm
université de technologie
Belfort-Montbéliard

RÉGION
BOURGOGNE
FRANCHE
COMTÉ

THÈSE DE DOCTORAT

DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE TECHNOLOGIE DE BELFORT-MONTBÉLIARD

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Informatique

par

YAZAN MUALLA

Explaining the Behavior of Remote Robots to Humans: An Agent-based Approach

Thèse présentée et soutenue à Belfort, le 30 Novembre 2020

Composition du Jury :

M. KOUKAM ABDERRAFIAA	Professeur à l'UTBM	Président
M. BALBO FLAVIEN	Professeur à l'École Nationale Supérieure des Mines de Saint-Étienne	Rapporteur
M. MATSON ERIC T.	Professeur à l'Université de Purdue, USA	Rapporteur
Mme GLEIZES MARIE-PIERRE	Professeur à l'Université Paul Sabatier de Toulouse	Examinatrice
M. VERCOUTER LAURENT	Professeur à l'INSA Rouen Normandie	Examinateur
M. GALLAND STÉPHANE	Professeur à l'UTBM	Directeur de thèse
M. NICOLLE CHRISTOPHE	Professeur à l'Université de Bourgogne	Codirecteur de thèse

N° 2020UBFCA023

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor *Prof. Stéphane Galland* who has been more than a supervisor to me, more like a mentor and a path creator throughout the journey. I have acquired much more knowledge and experience than I would have expected from a Ph.D. thanks to his endless support and inspiring guidance. I would also like to thank my co-supervisor *Prof. Christophe Nicolle* for his continuous motivation and worthwhile insights. My great appreciation to both.

I thank the respected jury members for their valuable evaluation and feedback. I thank also my supportive fellow researchers, colleagues, and co-authors who helped me in various aspects, namely: *Amro Najjar, Igor Tchappi Haman, Emmanuel Dimitry Ngounou Ntoukam, Timotheus Kampik, Cedric Paquet, Hui Zhao, Alaa Daoud, Davide Calvaresi, and Wenshuai Bai*. Additionally, I thank the members of the CIAD lab for the friendly and productive environment they have created.

This work is dedicated to everyone who helped me throughout this journey:

- My dear father *Ahmad* who is the eternal source of my strength, My mother's pure soul *Nuha* who stayed with me and lifted me throughout the years and will immortally stay with me, and My dear brothers *Alsamawal* & his family and *Safi* who are indispensable in my life ... I am forever indebted to you.
- My friends and family members, especially my family-in-law and my aunt *Jayda* ... thanks for the generous encouragement.
- My special dedication and gratitude go to my beloved wife *Salwa* who has the greatest credit in this achievement, my sweet little girl *Laure* for inspiring me with her beautiful presence and lovely laugh, and any precious son or daughter in the future. You are my hope and my source of pride and I hope to be the same to you.

  Salwa, I do not know if I am entitled to gift you with this, as the truth is you have gifted me with it and more. I cannot imagine achieving this without you my love . . . ever, and If life is fair my darling, you should deservedly acquire the title of doctor as well.

This journey was one of the most exhausting journeys of my life, but with you all being part of it, you made it easier and more endurable and, in the end, it was absolutely rewarding and totally worth it.

> To everyone who supported me in life . . . to my family, my wife, my daughter, and any other children in the future . . . I say I am proud of you, and I hope you are proud of me too.

Funding Statement

EN: This PhD thesis has received funding from the research program of the Council of Bourgogne Franche-Comté Region, France, under the UrbanFly project grant agreement 20174-06234/06242 (2017–2020).

FR: Cette thèse de Doctorat a été financée par le programme de recherche du Conseil Régional de la Région Bourgogne Franche-Comté (France), dans le cadre du projet UrbanFly 20174-06234/06242 (2017–2020).

CONTENTS

I

Background and Problem

1

INTRODUCTION

1.1/ CONTEXT

1.1.1/ EXPLAINABLE ARTIFICIAL INTELLIGENCE

Explaining the reasoning and the outcomes of complex computer programs has received considerable attention since the 1990s when research works on explainable expert systems were disseminated [272]. Nowadays, with the pervasive applications of machine learning, the need of explaining the reasoning of Artificial Intelligence (AI) is considered a top priority. Explainability helps in creating a suite of machine learning techniques that: *(i)* Produce more explainable models, while maintaining a high level of learning performance (*e.g.* prediction accuracy); *(ii)* Enable human users to understand, trust, and effectively manage the emerging generation of artificially intelligent entities [116]. 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 [38]. In April 2019, the European Union High-Level Expert Group on AI presented the Ethics Guidelines for Trustworthy AI [130]. 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 [77, 176, 233, 244]. Consequently, the sub-domain research of eXplainable Artificial Intelligence (XAI) gained momentum both in academia and industry [113, 16, 51]. Primarily, this surge is explained by the often useful, yet sometimes intriguing [275], results of black-box machine learning algorithms and the consequent need to understand how these data, fed into the algorithm, produced the given results [116, 34, 247]. An example of such intriguing results is when a Deep Neural Network (DNN) [322] mistakenly labels a tomato as a dog [275]. The aim is to interpret or provide meaning for an obscure machine learning model whose inner-workings are otherwise unknown or non-understandable by the human observer [16]. Another line of XAI research aims at explaining the outcomes of goal-

driven systems (*e.g.* robots) [16] since in the absence of a proper explanation, the human user will come up with an explanation that might be flawed or erroneous. This problem will be aggravated in 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, socially assistant robots, *etc*.). Ensuring mutual understandability among humans and robots becomes key to improve the acceptability of robots by humans and the human-robot interaction capabilities, and in particular to guarantee human safety in human-robot collaboration. The problem of understanding the behavior of robots is more accentuated in the case of remote robots since —as confirmed by recent studies in the literature [124, 23]— remote robots tend to instill less trust than robots that are co-located.

Despite considerable advances, the domain of XAI is still in its early stages of development, and to achieve smooth human-robot interaction and deliver the best possible explanation to the human, two key features have been outlined in the literature when providing an explanation [62, 222, 70]:

- **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 [273]. This becomes a challenge in complex situations involving multiple remote robots since this places more pressure on the human's cognitive load and requires adaptive XAI mechanisms able to cope with the limited human cognitive capabilities.

- **Adequacy**: refers to the need to include all the pertinent information in an explanation to help the human understand the situation. 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, this requires a specific explanation.

Recently, works in the literature have started to respond to these two features. In particular, the filtering of explanations has been suggested to achieve simplicity [120, 160]. Yet, the solutions offered in these works were not flexible enough to consider complex situations, *i.e.* the proposed models and experiments were not adaptive to changes in the environment and rather defined particularly for a specific situation. Additionally, there was a lack of determining the best granularity level (either detailed or abstract) of the explanation to avoid overwhelming the humans in various situations. A very recent work investigated different levels (none, detailed, and abstract) of explanations [185]. Their results showed that their model of abstract causal explanations provides better performance in terms of some explanation quality metrics but not in all, namely the "understand" metric for example. Moreover, the authors note that when comparing their model of explanation in a scenario with the same scenario but with no explanation, the results show no signifi-

cance for most of the explanation quality metrics.

To achieve adequacy, several works [145, 60, 202, 241, 302] investigated contrastive explanations, firstly pinpointed by Lipton [175], to provide explanations containing the necessary information needed by the human. This choice is supported by evidence from social sciences suggesting that, instead of providing a full explanation of the system, contrastive explanations can be more adequate, especially in abnormal situations [198]. Nevertheless, most of the works in the literature are carried out at the conceptual level with rare empirical human studies [198]. Moreover, some works [172, 96] considered contrastive questions like "Why didn't you do ...?", but not contrastive explanations.

One work by Kulesza et al. [162] tried to combine the two features and investigated the "sweet spot" between simplicity and adequacy. After a human study with only 17 participants, the result surprisingly showed that there is no sweet spot and that the solution is simply to give all the explanations possible to the human. One possible reason for such a result is the chosen settings of the experiment, as there was no challenging situation that provides too many explanations to overwhelm the human user; *i.e.* the work did not consider the human cognitive load.

In the context of human-robot collaboration, intelligent agents have been established as a suitable technique for implementing autonomous high-level control and decision-making in complex AI systems [318]. Agents are frequently applied to equip robots with greater autonomy. By designing proactive agents that control the robots, the latter become capable of autonomously managing their actions and behavior to reach their goals [17, 225, 207]. Some researchers have considered agents and robots to have indistinguishable roles in the AI system. For instance, Matson and Min [189] have introduced Human-Agent-Robot-Machine-Sensor (HARMS) model for interactions among heterogeneous actors. HARMS connects actors such that all of them are indistinguishable in terms of which type of actor (*e.g.* robot, software agent, or even human) sends a message [190].

An agent is defined as an autonomous software entity that is situated in some environment and where it is capable of actions and coordination with other agents to achieve specific goals [317]. For these reasons, the resulting Multi-agent System (MAS) technology has been established as a suitable platform for implementing autonomous behavior and decision-making in computer systems [309, 274, 318, 89]. Recent works on XAI for intelligent agents and MAS employ automatically generated folk psychology-based explanations [119, 45, 122]. These explanations communicate the beliefs and goals that led to the agent's behavior.

Our choice of an agent-based architecture is based on the characteristics of agents such as autonomy, responsiveness, distribution, and openness [318, 137, 89, 317, 309]. More specifically, we consider that agents are autonomous goal-driven entities that are bound

to an individual perspective. Thus, agents are capable of both representing the remote robot's perspective and piloting its interaction with its environment. Agents are also able to represent the human user and to apply his/her preferences regarding the interaction with the system and assess the explanations that he/she needs. Thus, the proposed Human-Agent Explainability Architecture (HAExA) is an agent-based architecture involving agents that represent robots handling the interaction with the human user. Through interaction and coordination of these agents, HAExA formulates explanations by relying on a mechanism of adaptive explanation filtering in conjunction with the use of contrastive explanations.

This thesis is in part of the research project UrbanFly 20174-06234/06242, supported by the Council of the "Bourgogne Franche-Comté" Region (France). One of the goals of this project is to propose novel models for simulating UAVs in urban environments and smart cities. In this context, the UAVs represent the remote robots explaining their behavior and actions to the human. Working with remote robots is a challenging task, especially in high-stakes and dynamic scenarios such as flying UAVs in urban environments. The next section investigates this challenge and outlines our response plan. It introduces the application of this thesis related to the explainability of UAVs.

1.1.2/ Artificial Intelligence in the Domain of Unmanned Aerial Vehicles

Since the early days of the industrial revolution, people started migrating to cities in droves. In 2007, the percentage of the urban population exceeded that of the rural population for the first time in history [282]. According to the United Nations, this urbanization is expected to accelerate raising the percentage of people living in cities and metropolitan areas to 68% of the world population by 2050 [281]. The result is a denser city infrastructure – where the city is the backdrop for all of the social, economic, and commercial activities. To accommodate this evolution, cities need to rely on technologies to help them improve the quality of life of their citizens.

UAVs, most commonly known as drones, are becoming increasingly popular for civilian applications in several domains such as agriculture, transportation, product delivery, energy, emergency response, telecommunication, environment preservation, and infrastructure. According to Teal Group's 2018 World civilian UAV Market Profile and Forecast report [280], civilian UAV production will total US$88.3 billion in the next decade, with a 12.9% compound annual growth rate. The same report states that the civilian UAV sector promises to be the most dynamic and growing sector of the world aerospace industry in the following years. Furthermore, fueled by growing demand from governments and private consumers, the civilian UAV market is expected to quadruplicate over the next

decade.

Currently, a new era of civilian UAVs that can autonomously fly outdoor and indoor is emerging. The key features making the UAVs interesting to use are their small dimensions, ability to take-off and land vertically, good maneuverability, simple mechanics, and payload capability. These features make UAVs accessible for civilian applications deployed in an urban environment where they started to be used as a practical solution for cost-efficient and rapid delivery. One of the most known examples is Amazon Prime Air where UAVs are used to deliver packages to customers [24, 72]. The future tendency is that UAVs will become more and more used, as new civilian applications are developing in people's daily life in urban environments.

Another important example that comes from the health-care sector is the transportation of medical samples and products, as immediately after being collected, the medical samples need to be conveyed to the sites of testing which can be located at a very long distance from the collection site [174]. To achieve that, a UAV health-care delivery network is used [253]. The main goal of this network is to facilitate more timely-efficient and economical drone healthcare delivery to potentially save lives. Even though there were some concerns about the safety of this procedure in such transportation environments, it has been shown that the UAV transportation systems are a viable option for the transportation of medical samples and products [13].

Despite these initial successes, UAV technology is still in its early stages of development. For this reason, considerable limitations should be addressed before a large scale deployment of UAVs in civilian applications is possible. One of the main limitations to mention is related to the high amount of energy consumed by these devices when staying airborne coupled with their limited battery life [242]. Moreover, since civilian applications are mostly deployed in urban environments involving multiple actors, considerable research efforts should be dedicated to enhancing the UAV perceptual intelligence required to coordinate complex environments [94], and more importantly to address the possible consequences, especially on people safety, of a mechanical failure that may cause a crash and the costs of such incidents [312].

To guarantee it is safe for UAVs to fly over people's heads and to reduce costs, different scenarios must be modeled and tested. However, currently, most of the regulations in force restrict the use of UAVs in cities. Even though some regulations were passed to regulate the use of UAVs [291] including air traffic, landing/taking off, *etc*., they are still immature and not yet fully developed. Moreover, legislation varies from region to region and between countries [57], and no proposals were made from a technological point of view. For this reason, to perform tests with real UAVs, one needs access to expensive hardware and field tests that are costly, time-consuming, and require trained and skilled people to pilot and maintain the UAVs. Moreover, in the field, it may also be difficult to

reproduce the same scenario several times [179]. To overcome these limitations, simulation frameworks have been developed to allow transferring real-world scenarios into executable models (*i.e.* simulating UAV activities in a digital environment) [205].

In multi-actor environments such as smart cities, there is a huge number of actors with a high density within the environment. Considering that a part of these actors will interact with intelligent, autonomous, and connected objects, or will be one of them, *e.g.* UAVs, it will be impossible to have a human associated with each of these objects, *e.g.* as a pilot. This implies a need for autonomy of these objects, cooperation among them for reaching their goals, and negotiation between them to avoid conflicts. The agent paradigm is well suitable for modeling, implementing, and deploying autonomous entities into multi-actor environments [106, 225]. Therefore, agents play a significant role in the coordination, cooperation, competition, and negotiation between all actors.

An Agent-based Simulation (ABS) simulates a model of a system that is comprised of individual, autonomous, and interacting agents that offers ways to more easily model individual behaviors and how they affect others in ways that have not been available before [182]. The results make ABS a natural step forward into understanding and managing the complexity of today's business and social systems. The use of ABS frameworks for UAVs is gaining more interest in complex civilian application scenarios where coordination and cooperation are necessary, *e.g.* the study of the swarms' formation of multiple UAVs [47, 248]. Despite these promising research efforts, very few works were dedicated to understand and analyze existing works using ABS in civilian UAV applications. Very few surveys outlined a comprehensive set of research questions pertaining to multi-agent simulations for civilian UAV applications.

1.2/ CONTRIBUTIONS OF THE THESIS

Explaining the behavior of robots is gaining more interest in the domain of Human-Computer Interaction (HCI) and particularly in the sub-domain of human-robot interaction, and this is a more challenging task in the case of remote robots, *e.g.* UAVs, explaining their behavior to the human. In this context, considerable merits are provided by agents when representing remote robots. More recently, XAI approaches have been extended to explain the complex behavior of goal-driven systems such as robots and agents. This thesis argues that enforcing measures of parsimony of explanation helps to meet the two desired features of an explanation, namely simplicity and adequacy. We qualify an explanation as parsimonious if *(i)* it adaptively, according to the complexity of the situation, filters the information provided to the human in a way that prevents overwhelming him/her with too many details; *(ii)* it relies on contrastive explanations to explain abnormal situations. To produce parsimonious explanations, the thesis introduces the process

of explanation formulation and proposes an architecture allowing to make this process operational.

Therefore, the general problem of this thesis is:

GENERAL PROBLEM

How to build an adaptive context-aware architecture, model, explanation process, and simulation tool to support the human-agent explainability for goal-driven AI systems in the context of remote robots (*e.g.* UAVs)?

Accordingly, two main objectives of the thesis could be emphasized:

Main Objective 1

Increase the understandability, hence the humans' confidence, of the behavior of remote robots through explainability.

Main Objective 2

Provide an architecture, a model, a process, and simulation tools for reproducing the complexity of the behavior of remote robots and the collective behavior.

More specifically, the theoretical contributions of the thesis are threefold:

1. Propose HAExA, an agent-based architecture that facilitates the human-agent explainability representing remote robots as agents. The architecture helps in formulating the necessary explanations communicated from remote agents to the humans, while at the same time considering the human cognitive load to avoid overwhelming him/her with too many details in the explanation.

2. Propose an *adaptive* and *context-aware* process of explanation formulation using various combinations of generating and communicating the explanations. We rely on generating and communicating the explanations on folk psychology, namely the Belief-Desire-Intention (BDI) model [238], as it helps in generating context-aware explanations, and in adaptively communicating the explanations. This leads us to explore the concept of parsimony of explanations that could help in simplifying the explanations with different explanation communication techniques like the filtering of explanations while keeping all the necessary information, especially in abnormal situations where contrastive explanations are used.

3. Develop ABS tools[1] to implement a proof of concept of the proposed model and architecture. These tools are built based on HCI capabilities to facilitate the subjective evaluation of the explanation approaches in the proposal by humans participating in the evaluation process.

Miller [198] has addressed the lack of empirical human studies in the domain of XAI as a shortage in the literature. Therefore, and in addition to the three mentioned theoretical contributions, there is one contribution related to experimental validation. It is to conduct two empirical human case studies based on a scenario of package delivery using civilian UAVs: First, a pilot test, investigating the role of filtering of explanations in three cases (No explanation, detailed explanation, and filtered explanation). Second, the main test, investigating different techniques of explanation formulation (static filter, adaptive filter, and adaptive filter with contrastive explanations). The significance of the participants' responses is statistically analyzed and presented using non-parametric and parametric testing.

1.3/ Outline of the Thesis

The thesis is structured in five parts that are described below. Figure 1.1 depicts the different contributions of this thesis in relation to the structure of the thesis.

Part I (Context and Problem): Apart from this chapter that provides the introduction, Chapter 2 lays down the background concepts and definitions.

Part II (State of the Art): Within this part, Chapter 3 analyses the most related works in the XAI literature, while Chapter 4 provides a Systematic Literature Review (SLR) of ABS in the domain of UAVs.

Part III (Contribution): It provides the theoretical contributions of the thesis. Chapter 5 aims at positioning the thesis through a research methodology. The methodology recalls the objectives of the thesis and highlights the related research questions and hypotheses. Chapter 6 proposes HAExA, the human-agent explainability architecture and thoroughly discusses the explanation formulation process.

Part IV (Evaluation): It aims to thoroughly evaluate the contribution of the thesis. Chapter 7 presents the empirical case study and the questionnaire built to collect the responses

[1] As part of the UrbanFly project.

of the human participants. Chapter 8 and Chapter 9 perform, respectively, the pilot test and the main test conducted employing different ABS tools. In each of these two chapters, the responses of the participants are statistically analyzed and the results are validated, presented, and investigated.

Part V (Conclusion and Perspectives): It concludes this thesis with two chapters. Chapter 10 concludes the thesis with a summary and a general discussion and provides future perspectives.

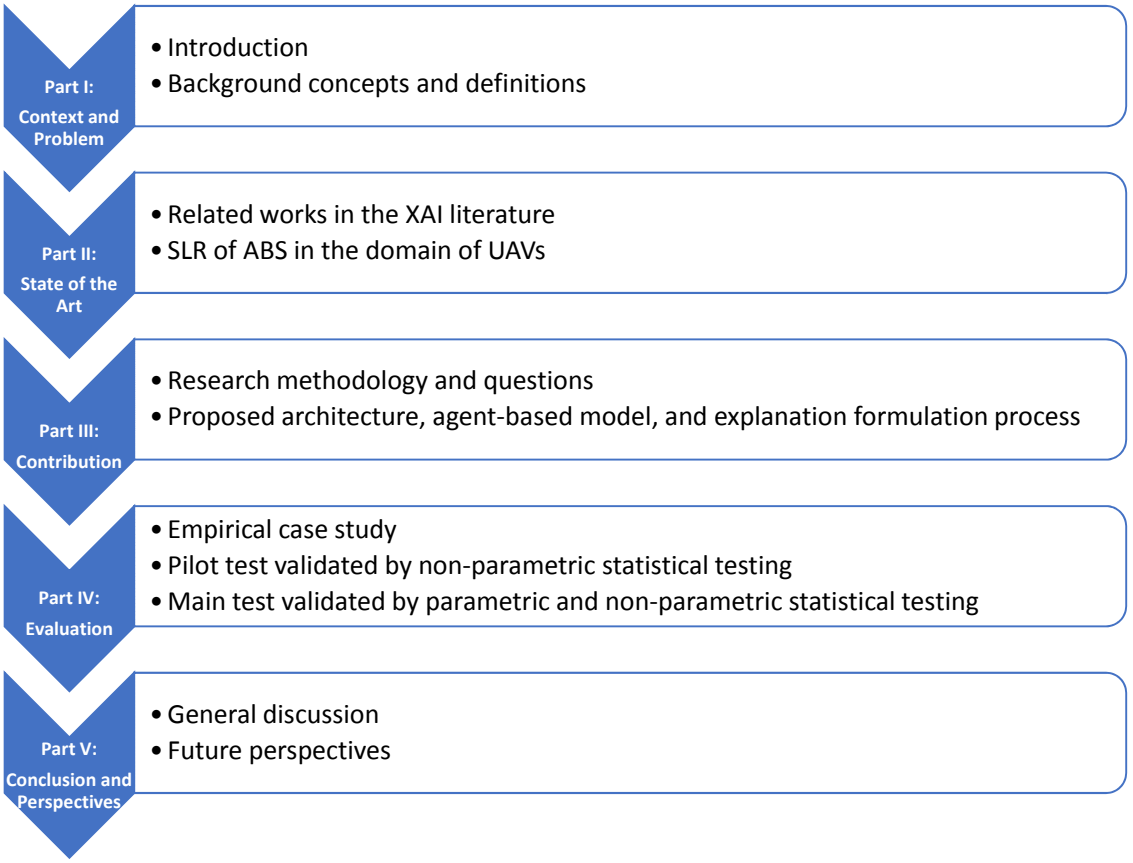


Figure 1.1: Outline of the thesis

2

DEFINITIONS

2.1/ INTRODUCTION

This chapter sketches the background definitions and fundamental concepts of this thesis. First, Section 2.2 defines the concept of remote robots, and Section 2.3 details the features allowing agents to be a good representative of remote robots. Section 2.4 offers a brief introduction to the goal-driven explainable agents and robots. Then, Section 2.5 introduces the concept of parsimony and discusses its relation to Explainable Artificial Intelligence (XAI), while Section 2.6 introduces the contrastive explanation as a component of a parsimonious explanation. Section 2.7 outlines the process of providing an explanation, while Section 2.8 discusses the cognitive architectures allowing to implement such a process. Section 2.9 explains the concept of agent-based modeling and simulation that will be used to implement the proposal and evaluate it in the empirical human studies. Finally, Section 2.10 discusses various ways to build XAI-based questionnaires to be used in the empirical human studies.

2.2/ REMOTE ROBOTS

In the last decades, the concept of robotics has largely involved remotely-operated mobile robots equipped with cameras being used to get eyes on something out of reach. However, autonomy in robotics means the ability of the robot to make its own decisions. There is no agreed definition of what is an autonomous robot, *a.k.a.* an autorobot, but this thesis adopts the definition provided in Definition 1. The main point we emphasize is that the crucial components of an autonomous robot are perception, decision making, and actuation.

Definition 1: Autonomous robot according to Bekey [29]

An intelligent machine capable of performing tasks in the world by itself, without explicit human control.

Remote autonomous robots, or remote robots for short, are a special type of autonomous robots, and several definitions are associated with remote robots. In this thesis, we adopt that a remote robot is a specific type of autonomous robots with the following two characteristics: *(i)* Autonomy: a remote robot must retain its state as it travels and conduct tasks autonomously in the environment to serve a purpose or goal. The robot acts to the change of the environment or its status and influences what it sensed. *(ii)* Mobility: a remote robot must be able to travel to remote destinations, primarily by self-determination, *i.e.* it is not co-located with the human. It is important to distinguish remote robots from remote-controlled robots that are controlled by the human user, *i.e.* not autonomous.

Some researchers have considered agents and robots to have indistinguishable roles in the Artificial Intelligence (AI) system. For instance, Matson and Min [189] have introduced Human-Agent-Robot-Machine-Sensor (HARMS) multi-agent system model for interactions among heterogeneous actors. HARMS focuses more on accomplishing a task rather than specifically which agent really carries out the task. HARMS communication model is used to indistinguishably facilitate communications between actors of the systems including agents and robots [190, 170].

Autonomous agents are frequently applied to equip autonomous robots, including remote ones, with greater autonomy. By designing proactive agents that control the robots, the latter become capable of autonomously managing their actions and behavior to reach their goals [17, 225, 207]. The next section provides the definitions related to autonomous agents.

2.3/ Intelligent Agents

Wooldridge and Jennings [318] have provided one of the most common definitions of intelligent agents (Definition 2).

Definition 2: Agent according to Wooldridge and Jennings [318]

A computer system situated in some environment, that is capable of autonomous action in this environment to meet its design objectives.

We argue that agents hold specific features that make them suitable to represent remote robots [318, 137, 89, 317, 309]:

Autonomous: Automation will minimize human intervention. As a result, it will save time

as robots are designed to work with no delay. Additionally, in a real-time application like the one addressed by this thesis, autonomy can reduce the delay introduced by human intervention. Moreover, in large scale situations, it is impractical to allocate each robot with a human to control it. Additionally, some aerial tasks are more efficient in resources to be executed by robots, *e.g.* Unmanned Aerial Vehicles (UAVs), than by humans.

Decentralized: Decentralization of the problem is justified by two reasons: *(i)* the physical distance between the remote robots and the human in one hand, and between the remote robots themselves. *(ii)* each robot has different specifications to complete the tasks. Therefore, computing an optimal centralized solution when there are such specifications is a computationally difficult task and takes time which is crucial in such real-time applications.

Reactive: Reactivity is an important requirement in real-time applications, and agents are reactive to what they sense. The specification of each robot is different and accordingly, it responds differently.

Cognitive: In this thesis, we rely on agents with cognitive architectures to represent the reasoning of remote robots. This helps in the interaction with the humans that have a cognitive manner of thinking. The remote robots collect the data from the environment, *i.e.* beliefs, and have an internal reasoning loop to allow them to make decisions to serve a specific goal.

Flexible (or open): Robots could be, arbitrary, added in the system, removed, and put back from a mission to the next mission. With this in mind, the use of agents is an efficient approach to achieve this requirement due to their intrinsic modularity. In other words, agents can be easily added or removed, without the need for detailed rewriting of the system. This feature also helps in preventing the propagation of faults, and in self-recovery. Additionally, as the system is flexible in terms of robots chosen for a mission, backup robots can be used instead of the failed ones already in a mission, which provides the system with a fault-tolerant feature.

Social: Robots represented as agents can communicate with each other and with the physical components. They can cooperate, collaborate, or even compete or negotiate. In terms of the overall benefit of the system, this helps in deciding which robot to conduct the mission or if there is a need for several robots to conduct the mission together, *e.g.* a swarm of UAVs lifting a heavy object. Even though robots are self-interested in terms of achieving their goals, the overall welfare of the system should be considered as all robots share the same main goal which is to serve the human. Moreover, each robot has different criteria, *e.g.* capacity load in case of UAVs lifting objects, to complete the tasks of the mission. Therefore, there is a need for a mechanism to reach a collective decision for the benefit of the whole system.

While agents have been established as a suitable technique for implementing autonomous high-level control and decision-making in complex AI systems [318], there is still a need for these systems to be understood and trusted by human users. The next section defines how explainability is used in the domain of goal-driven systems such as robots and agents.

2.4/ GOAL-DRIVEN EXPLAINABLE ARTIFICIAL INTELLIGENCE

According to our knowledge, a formal and universally agreed definition of XAI is lacking. However, the definition that is adopted in this thesis is provided in Definition 3.

Definition 3: Explainable Artificial Intelligence according to Adadi and Berrada [4]

A suite of techniques that produce more explainable models to make a shift towards more transparent AI.

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 Deep Neural Networks to take a given decision (*e.g.* a classification decision) [113]. Goal-driven XAI is defined as in Definition 4.

Definition 4: Goal-driven XAI according to Anjomshoae et al. [16]

A research domain aiming at building explainable agents and robots capable of explaining their behavior to a human.

More recently, XAI approaches have been extended to explain the complex behavior of goal-driven systems such as robots and agents [16, 127]. The main motivations for this move are:

(i) In general, human-robot interaction is a key challenge, since, by default, it is not straight-forward for humans to understand the robot's State-of-Mind (SoM) defined in Definition 5.

Definition 5: Robot State-of-Mind according to [127]

The non-physical entities of a robot such as intentions and goals.

As has been shown in the literature, humans tend to assume that these robots/agents have their own SoM [127], and that with the absence of a proper explanation, the human will come up with an explanation that might be flawed or erroneous;

(ii) In the near future, these goal-driven systems are expected to be omnipresent in our daily lives (*e.g.* social assisting agents and virtual assistants) [16].

Therefore, ensuring mutual understandability among humans and robots/agents is key to improve the acceptability of robots/agents by humans, and in particular to facilitate human safety in human-robot collaborations. However, humans have learned and developed a natural ability to understand others, and in human-robot interaction scenarios, this ability is still adopted even if very limited for now [127]. In this context, goal-driven XAI is of particular interest since providing explanations in multi-agent environments is even more challenging than providing explanations in other settings [19]. Led by that, the next section introduces the concept of parsimony of explanations.

2.5/ PARSIMONY AND EXPLAINABLE ARTIFICIAL INTELLIGENCE

The concept of parsimony of explanations has received considerable attention for centuries. A famous formulation of this concept is the "Occam's Razor" [284, 36] stipulating that: "Entities should not be multiplied beyond necessity." Thereafter, Occam's Razor[1] 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 [240, 110, 183].

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 shown in the literature [164, 315, 158], parsimony is a largely subjective quality. Therefore, and even if the human preference for simplicity is not necessarily true, we could assume that if two explanations are equal in theory, the one with the empirical evident support is the best. For this reason, empirical human studies have been outlined as key to assess how parsimonious an explanation is to a given user in a given situation. In these studies, the opinions of the participants 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 provided by the AI systems to their human users should be simple while containing all the pertinent information of the system's decision. Thus, parsimony has been identified as a key desideratum for XAI [239]. Yet, very few works in the literature are proposed to define what parsimony means in the context of XAI? how parsimonious explanations can be generated and communicated to the humans? and how their impact on the humans receiving them is assessed? (refer to Chapter 3 for an overview of these works). In this thesis, we forge and adopt a definition of a *parsimonious explanation* based on the "Occam's Razor" and the literature of XAI (Definition 6).

[1] William of Occam, 1290–1349.

Definition 6: Parsimonious explanation

The simplest explanation that includes all the necessary information for the human to understand the situation.

The discussion of the parsimony of explanations opens the door to the questions: *what information is necessary to be kept in an explanation?* and *how to formulate a parsimonious explanation?* To tackle these questions, some works investigated contrastive explanations as a potential way to generate explanations that include the necessary information that the human needs, instead of providing a full explanation of the system (*e.g.* [185]). The next section offers an overview of contrastive explanations.

2.6/ 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 [198]. The aim is to explore how humans generate and communicate explanations in their everyday life. Everyday explanations are explanations of why particular events, behaviors, decisions, *etc*. happened [187]. Evidence in the literature suggests that in abnormal situations, these everyday explanations should take the form of contrastive explanations [198].

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, they prefer selecting one or two causes from a sometimes-infinite number of causes to be the explanation. However, this selection is influenced by certain cognitive biases [198]. Lipton [175] has 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 [175]. Later research showed that people do not explain the causes for an event by itself, but they explain the cause of an event relative to another counterfactual event (that did not occur). Therefore, according to Kim et al. [145], 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 [129, 128]. This contrastive explanation may take the form of 'why' questions and be expressed in various ways [278, 169].

In recent years, research on contrastive explanations in AI received a growing attention [60, 202, 241, 302]. Lim and Dey [172] found that "Why not ...?" questions were common questions that people asked after some human studies on context-aware applications. The definition of context-aware explanations is provided in Definition 7.

Definition 7: Context-aware explanations according to Anjomshoae et al. [16]

Explanations where the agents/robots consider the context when explaining the situation.

Winikoff [314] investigated how to answer contrastive questions, *e.g.* "Why didn't you do ...?" for Belief–Desire–Intention (BDI) programs. Another similar work has checked a similar type of questions like "Why didn't you do something else" [96]. However, most of the existing work 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 [198].

Evidence from social sciences confirms the importance of contrastive explanations both in human-to-human explanations and computer-to-human explanations. In an influential recent survey, Miller [198] has identified useful insights related to XAI from social sciences. Among the other key findings outlined in his work, he postulated that explanations are contrastive in the sense that they are responses to particular counterfactual cases. The next section expresses the steps of providing an explanation by agents to the human.

2.7/ PHASES OF AN EXPLANATION

Neerincx et al. [214] have emphasized the fact that for the explanations to serve their purposes they should be aware of the context of the environment and the human information processing capabilities, *i.e.* human cognitive load. The latter is defined in Definition 8.

Definition 8: Human cognitive load according to Sweller [273]

A limit beyond which humans are unable to process the provided information.

According to Neerincx et al. [214], the process of providing explanations by agents to the human includes three distinct phases:

- **Generation:** This phase considers what to explain to the human. For example, explaining the perceptual foundation of the agent behavior, or explaining why a certain action is applied. The aim is to generate an explanation justifying why an action was taken. The actual implementation of this phase is determined by the agent model (*e.g.* BDI agent [238]). Citing goals [45], desires [141], and emotions [142] are examples of the explanation generation process in the literature.
- **Communication:** This phase is about the form of the explanation (textual, visual, in a simulation, *etc.*) and the means to communicate the explanation (*e.g.* Knowledge

Query and Manipulation Language [91]). this phase deals with how to provide and present the explanation to the end-user [214].

- **Reception:** This phase investigates how well the human understands the explanation. To assess this, research relies on human studies and subjective evaluation. Furthermore, to better understand the explanation reception, meaningful metrics should be devised to assess the explanation and poll the users about it [16]. Concerning XAI reception, some user studies (*e.g.* [212]) have been conducted, but there is a significant lack of empirical studies involving human users in realistic human-agent settings and scenarios where explanations are needed to understand the system's behavior [198, 199].

To facilitate these three phases into one complete process and in particular to empower the agent with the ability to build the explanations properly, the next section explains one approach used for supporting the explanation process modeling and implementation.

2.8/ Cognitive Agent Architectures for Supporting the Explanation Process

Cognitive agent architectures in the applications related to explainability are gaining more interest lately [16]. Any proposed architecture should have the following characteristics:

(i) A representation of the environment where the agents act and interact;

(ii) A self-representation of the agent's internal reasoning cycle;

(iii) Social skills for interacting with other agents.

These characteristics can be found —to different extents— in several well-known cognitive architectures such as BDI [41], FORR [84], ACT-R [14], LIDA [97], or Soar [165]. All these architectures reflect the first and second previously mentioned characteristics. Soar, BDI, ACT-R, and CLARION allow, additionally, to create social agents. ACT-R and CLARION [270] architectures are time-consuming to compute; hence they are not scalable in the context of near-real-time applications. The BDI architecture allows agents to exhibit more complex behavior than purely reactive architectures but without the computational overhead of other cognitive architectures [5]. Moreover, some evidence exists that BDI agent architectures facilitate knowledge elicitation from domain experts [85]. Furthermore, because BDI is based on the concepts of folk-psychology, it has been outlined as a good candidate to represent everyday explanations [217, 45] since it is considered as the attribution of human behavior using *'everyday'* terms such as beliefs, desires, intentions, emotions, and personality traits [63, 186]. Folk psychology [63] refers to the

explanation of human behavior in terms of his/her underlying mental states such as beliefs, desires, and intentions. Folk psychology is how humans in everyday communication explain and predict intentional actions [187, 63]. It means that "the core concepts of the agent framework map easily to the language people use to describe their reasoning and actions in everyday conversations" [217]. The idea of programming computer systems in terms of *mentalistic* notions like beliefs, desires, and intentions is a key component of the BDI model. The concept was first articulated by Yoav Shoham, in his Agent-oriented Programming (AOP) proposal [260].

Existing works in the literature present considerable advances in cognitive agent architectures. However, most of them do not support explainability functions. To further push the research of goal-driven XAI, linking the agent inner model with the explanation generation module is a crucial step [16].

For all the previously mentioned reasons, this thesis considers BDI architecture as a good option for providing contrastive explanations since it relies on folk-psychology to represent everyday explanations. Therefore, we opt to adopt it in the proposed architecture in this work (Chapter 6).

To evaluate the proposed agent-based architecture, an expected implementation tool is the simulation, and more specifically Agent-based Simulation (ABS). The next section provides the reasons to support this decision.

2.9/ AGENT-BASED MODELING AND SIMULATION

Due to the complexity exhibited by a complex system, a suitable approach is required to study, investigate, and implement it. Modeling and simulation are important candidates to achieve that [321]. For years, several modeling and simulation practices have been developed in several fields of science. Modeling and simulation allow to understand, predict, and even control real or virtual phenomena [320]. They enable the study of real or virtual phenomena in laboratories to produce related knowledge. Therefore, modeling and simulation are appropriate for studying systems that can not be directly observed or measured [105]. Modeling and simulation theory is not based directly on the system to be studied but on a simplification of it. The source system is the real or virtual environment to model that represents a part of the reality circumstances [287].

There is currently no interdisciplinary consensus on the definition of the term *model*. This thesis considers the definition stated in Definition 9. This definition highlights two of the main characteristics of a model: *(i)* Understand the functionalities of the source system that enables the prediction of its evolution. *(ii)* Answer the questions about the source system where the model captures only aspects of the source system needed to be

answered. Therefore, a model cannot be used in general to answer any question about the source system that it represents, but only to those for which it was designed.

Definition 9: Model according to Treuil et al. [287]

An abstract construction that makes it possible to understand the functioning of a source system by answering a question concerning it.

A simulation can be used to test a hypothesis of the source system, verify it, or accredit the theory that was used to build it. A simulation can be also used to understand the functionalities of the source system and therefore serve as a support for decision making [321].

Like the concept of a model, a simulation does not have a consensual definition in the literature. Ören [221] has collected more than 100 different definitions of a simulation [221] and about 400 different types of simulation [220]. We adopt in this thesis the definition of a simulation stipulated in Definition 10 and 11. Definition 10 allows to explain the link between a model and its simulation. A simulation consists of executing a model and changing its states step by step according to its dynamics while producing the outputs associated with each state. Therefore, a simulation generates information on the evolution of the model over time. Definition 11 states that the objective of the simulation is to imitate some real-world phenomena. Therefore, a simulation can be applied to a wide range of real-world phenomena.

Definition 10: Simulation according to Treuil et al. [287], Cellier and Greifeneder [54]

An experimentation performed on a model.

Definition 11: Simulation according to Law et al. [167]

A set of techniques that employ computers to imitate – or simulate – the operations of various kinds of real-world facilities or processes.

On the same vein, the definitions of agent-based modeling and ABS are provided in Definition 12 and Definition 13 respectively.

Definition 12: Agent-based modeling according to Macal and North [182]

Modeling systems comprised of individual, autonomous, and interacting agents that offers ways to more easily model individual behaviors and how they affect others in ways that have not been available before.

Definition 13: Agent-based simulation according to Macal and North [182]

An approach to simulate agent-based models.

Consequently, ABS can be considered as a natural step forward towards better understanding and managing the complexity of today's business and social systems. Additionally, cognitive architectures are frequently applied in ABS [5]. Some researchers refer sometimes to ABS as Multi-agent-based Simulation (MABS) that combines the advantages of the Multi-agent Systems (MAS) paradigm, with the advantages of the simulation [229, 22, 21]. The emphasis in MABS is more on the cooperation and collaboration of the agents.

For the reasons mentioned in this section and the previous one (Section 2.8), ABS of BDI agents is considered as a good candidate to simulate the behavior of complex agent-based systems thereby offering a platform to build explainable agents and assess their understandability from the human user perspectives. Thus, this thesis proposes an explainable BDI agent model 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 the human users.

To complete the evaluation of the proposal, human case studies are conducted. Therefore, there is a need for an appropriate manner to collect and aggregate the responses of the participants. The next section discusses how questionnaires are built in the domain of XAI.

2.10/ Explainable Artificial Intelligence Questionnaires

There are several methods and related questionnaires 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 [132]. 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. In the following, we state two ways to build the range of the questions and answers of the questionnaire [132].

2.10.1/ Explanation Quality Checklist

One way to build the questions is the Explanation Quality Checklist[2], which can be used by XAI researchers to either design high-quality explanations into their system or evalu-

[2]named "Explanation Goodness Checklist" in [132].

ate the quality of the explanations of a given system. In this checklist, only two choices (Yes/No) are provided. The goal of this checklist is to confirm aspects like understandability, satisfaction, level of details, completeness, reliability, trustworthiness, *etc*. However, this binary scale does not allow for being neutral. Besides, for some aspects, there is a need for more granularity, *i.e.* the use of more options of the results.

2.10.2/ Explanation Satisfaction and Trust Scale

This scale measures both trust and satisfaction (or understandability). This scale is recommended in the XAI domain, as it is based on the literature in cognitive psychology, philosophy of science, and other pertinent disciplines regarding the features that make explanations good [132]. Studies showed that with the scale of three answers, the participants tend, usually, to choose the middle option because they prefer not to be extremist in their choices. Therefore, in social science, the scale of answers is usually distributed to 5 answers allowing the participants to step away from the middle without the feeling of being in the extreme edges. In this context, the Likert scale [10] 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 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 validation part of this thesis (*see* Section 7.3)[3].

2.11/ Conclusion

Agents have been established as a suitable technique for implementing autonomous high-level control and decision-making for goal-driven systems like remote robots. This is thanks to their features: autonomous, decentralized, reactive, cognitive, flexible, and social. However, there is still a need for these goal-driven systems to be understood and trusted by human users.

With the advent of XAI, research on parsimony of explanations has gained new momentum since the explanations provided by the goal-driven systems to their human users should be simple while containing all the pertinent information related to the system's decision. Thus, the parsimony of explanations has been identified as a key desideratum for XAI. The aim is to explore how humans generate, communicate, and receive explanations in their everyday life. As the BDI model is based on the concepts of folk-psychology, it has been outlined as a good candidate to represent everyday explanations, hence represent the parsimony of explanations.

[3] A choice validated by other relevant works in the literature [132].

After proposing an agent-based model and architecture to facilitate and formulate the explanations, there is a need to evaluate the proposal. ABS has been established as a good candidate to simulate the proposal as a complex system thereby offering a platform to build explainable agents and assess their understandability from the human user perspectives.

In the end and as parsimony is a largely subjective quality, human studies have been outlined as key to assess how parsimonious an explanation is to a given user in a given situation. In these studies, the opinions of humans on the usefulness of explanations are collected and analyzed, and for that, there is a need to build appropriate questionnaires that are based on the recommendations and metrics from the XAI domain. In particular, the explanation satisfaction and trust scale is preferably used in such a context.

After providing the required definitions and background concepts in this chapter, the next two chapters discuss in detail the state of the art (*i.e.* positioning this thesis facing the existing works).

II

State of the Art

3

Explainability and Explainable Artificial Intelligence

3.1/ Introduction

As humans increasingly depend on complex AI systems, it becomes increasingly important to provide explanations for the decisions of these systems to foster effective human-system interaction. This orientation is confirmed by the ratification of the recent law on the General Data Protection Regulation (GDPR), which underlines the right to explanations [52]. Therefore, the design of transparent and intelligible technologies becomes a pressing necessity [16]. Recently, the Explainable Artificial Intelligence (XAI) domain has emerged to promote transparency and trustworthiness. Several reviews about the works addressing XAI were provided, but most of them deal with data-driven XAI to overcome the opacity of black-box algorithms, while few reviews were conducted to investigate the goal-driven XAI (*e.g.* an explainable agency for robots and agents).

The main three goals of this chapter are to: *(i)* understand the context of the thesis by investigating the contributions in the goal-driven XAI domain *(ii)* understand how the literature approached the realization of the explanations in terms of the three phases of generation, communication, and reception. *(iii)* outline the open research issues in this domain that will be the basis for the research questions and hypotheses of this thesis. These three goals will generally help us in adopting what is useful in each phase of providing an explanation when proposing our architecture, model, and process of explanation formulation.

The rest of this chapter is organized as follows. Section 3.2 sheds the light on the contributions of the goal-driven XAI and Section 3.3 explores the related works about the parsimony of explanations by investigating the key features outlined in the literature when providing an explanation. Finally, Section 3.4 concludes this chapter linking it with the rest of the chapters and sections in this thesis.

3.2/ Related Work on Goal-driven Explainable Artificial Intelligence

3.2.1/ Background and Origin

Research on XAI [116] has grown tremendously in recent years. However, recent studies on explanatory methods have mainly focused on data-driven algorithms aimed at interpreting the results of black-box machine learning mechanisms such as Deep Neural Networks (DNNs) [322]. This kind of research, driven by the intriguing results of DNNs (*e.g.* a DNN erroneously labeling a tomato as a dog [275]), intends to interpret or make sense of an obscure machine learning model whose inner mechanisms are otherwise unknown or incomprehensible to the humans. Thus, the majority of these studies focus on providing an overview of the explainability of data-driven algorithms [34, 79, 199, 113, 247] despite that agents are becoming ubiquitous in the daily life of humans in several applications.

Studies in the XAI literature have underlined the importance of taking into account the intended objective when incorporating means of explanation in intelligent systems [126]. Increasing human confidence in the system, transparency (*i.e.* explaining the internal workings of systems to the human), and informing about the agent's intentions (communication of intentions) are among the main motivations behind the explanations in the literature [16]. Studies suggest that transparency and trust go hand in hand to increase the human confidence in the system by understanding how its reasoning mechanism works (*e.g.* [303, 55]). In applications requiring human-agent interaction, intention communication is one of the main explanatory drivers to make the agent's internal state (*e.g.* goals and intentions) understandable for humans [25].

Goal-driven XAI (*see* Section 2.4) aims at building explainable agents and robots capable of explaining their behavior to humans [16]. These explanations help humans better understand the agent and lead to better human-agent interaction. They would encourage the human to understand the capabilities and limitations of agents, thus improving trust and safety levels, and avoiding failures, as the lack of appropriate mental models and knowledge about the agent can lead to failed interactions [58, 33]. The next section discusses first a review addressing goal-driven XAI and second the main drawbacks of the works in the literature of the domain of goal-driven XAI.

3.2.2/ Review on Goal-driven Explainable Artificial Intelligence

A very recent Systematic Literature Review (SLR)[1] has been conducted targeting contributions addressing goal-driven XAI [16]. The objective of this SLR was to clarify, map,

[1] *see* Definition 14 on page 42.

and analyze the relevant literature on explainable agents and robots over the last ten years (2009-2019). The SLR included 62 papers after a coarse-grained and fine-grained examination. The chronological distribution of work on the goal-driven XAI shows an increasing growth over the last five years [16]. This could be due to the effect of the general emphasis on explainable AI and the right to explanation by the GDPR [298] and similar initiatives [52].

From the same SLR [16], a research question explored the literature to find out whether the studied works were based on a background in social science or psychology. 39 of the 62 works did not have any theoretical foundation linked to the generation of explanations. For the rest, the most cited social science theory has been shown to be folk psychology (*see* Section 2.8). Additionally, another research question focused on the platforms and architectures used to design the goal-driven XAI. Besides ad hoc solutions, the most used architecture is the Belief–Desire–Intention (BDI) architecture (refer to Section 2.8 for more details on this architecture) to generate explanations for goal-driven agents (*e.g.* [45, 214]) [16]. The same recommendation is supported in another very recent review on goal-driven XAI [246].

The same SLR also found that context-aware explanations (*see* Definition 7 on page 19) have been proposed to implement effective control in ubiquitous systems (*e.g.* [173]), to facilitate context-aware applications in human-robot collaboration (*e.g.* [109]), and to improve robot navigation (*e.g.* [95, 123]).

Considering the types and categories of explanations. The SLR concluded that the most common type is that of textual explanations [16]. Textual explanations are sometimes presented in the form of natural language processing (*e.g.* [303, 121]). Regarding categories, the SLR has shown that the most important category is introspective informative explanations. This type of explanation is based on the reasoning process that leads to a decision to improve the quality of the interaction between humans and agents [16]. It is worth mentioning here that for several studies, the category of explanation was the contrastive explanation (*see* Section 2.6), *e.g.* [214, 56, 268].

The main drawback of the works in the literature of the domain of goal-driven XAI is the lack of empirical evaluations or conducting a user study for relatively simple scenarios [16, 198]. Other drawbacks or challenges include the communication of the explanations, issues related to the core AI running the system, and context-awareness [16]. In this thesis, we tackle these challenges and drawbacks both in the contribution and the evaluation parts. The next section gives an overview of the related work to our approach of parsimony of explanations.

3.3/ Parsimony in Explainable Artificial Intelligence

3.3.1/ Explanation Features

The domain of XAI is still in its early stages of development, and 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 [62, 222, 70]:

- **Simplicity**: providing a relatively simple explanation that considers the human cognitive load (*see* Section 2.7). This becomes a challenge in complex situations involving multiple remote agents since this places more pressure on the human's cognitive load and requires adaptive XAI mechanisms able to cope with the limited human cognitive capabilities.

- **Adequacy**: refers to the need to include all the pertinent information in an explanation to help the human understand the situation. Adequacy turns out to be a challenge in abnormal situations, where the agent tends to diverge from the behavior expected by their human users, and therefore, this requires a specific explanation.

Recently, works in the literature have started to respond to these two key features. The next sections investigate these works in detail along with their approaches to adhere to the explanation features.

3.3.1.1/ Addressing Simplicity

Previous works on XAI showed how to use the beliefs and goals of an agent for generating action explanations [122, 44]. However, this means that designing and formulating the beliefs and goals of the agent are necessary [122, 279].

Two common explanations styles in folk psychology are goal-based and belief-based explanations [75, 187, 63, 186]. A goal-based explanation communicates the agent goals (*i.e.* intentions) as the outcome of the action. It answers the question: 'For what purpose?' A belief-based explanation provides information on why the agent chose a specific action to execute. It offers information about the context and the circumstances. In this type of model [119, 141, 142, 45], the generation of explanation is based on the concept of hierarchical task analysis [250]. The latter is a well-established technique in cognitive task analysis that connects internal reasoning processes to external actions using Goal Hierarchy Trees (GHT) [250]. The latter has one main task divided into sub-tasks which are divided into sub-tasks, *etc*. Sub-tasks that are not divided represent the actions that can be directly executed in the environment. This tree includes tasks and adoption conditions

in a task hierarchy that can be seen as goals and beliefs. The explanation in the tree could either be based on the goals (*i.e.* intentions) or the beliefs of the cognitive agent.

Harbers et al. [121] have introduced a theoretical framework for explaining agent behavior, and some guidelines for developing explainable cognitive BDI models [122, 119]. The generation of explanations in their work is based on the concept of hierarchical task analysis. In our work, we use the same concept for generating some types of normal explanations (*see* Section 6.5.1.1 on page 93). However, we extend this concept to generate contrastive explanations (*see* Section 6.5.1.2 on page 95).

An interesting example of the work of Harbers et al. [121] has discussed the generation and the granularity (either detailed or abstract) of the explanation with a firefighting application [120]. However, the work is not conclusive in preferring a granularity level. Moreover, and while the work 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 detail, in this special case, is not considered; *i.e.* the work does not identify a threshold level, beyond which explanations are overwhelming for humans.

One work by Kulesza et al. [160] has evaluated the amount of information provided to the human and how it affects the understandability [160]. This work has investigated what a human user needs to know to productively work with an agent. The domain of the work is related to expert systems, and the goal is to investigate if human users can understand how these systems operate to fix their agent's personalized behavior. The model of these authors seeks to allow humans to build a mental model of the agent's reasoning. Mental models of agents were first discussed by Tullio et al. [289]. Previous work has found that humans may change their mental models of an agent when the agent makes its reasoning transparent [161]. However, some explanations by agents may lead to only shallow mental models [267]. Therefore, according to Kulesza et al. [160], making the agents' reasoning more transparent to the human is one way to influence the mental models. Kulesza et al. [160] have explored the effects of mental model soundness on the personalization of the agent by providing structural knowledge of a music recommender system in an empirical human study. The results of the study show that providing humans with detailed explanations about an agent's reasoning can increase their understanding of how the system works. However, information comes at the price of attention, as the human's time and interest are finite, so the solution may not simply be "the more information, the better" [160]. In our thesis, 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.

3.3.1.2/ Addressing Adequacy

To achieve adequacy, several works [145, 60, 202, 241, 302, 214, 56, 268] investigated contrastive explanations, firstly pinpointed by Lipton [175], to offer explanations containing the necessary information needed by the human (refer to Section 2.6 for more details on contrastive explanations). This choice is supported by evidence from social sciences suggesting that, instead of providing a full explanation of the system, contrastive explanations can be more adequate, especially in abnormal situations [198]. Nevertheless, the models used in these works have a limited adaptation to changes in the environment, *i.e.* they simply included a reactive behavior to events in the environment when building the contrastive explanation. Moreover, most of the works in the literature are carried out at the conceptual level with rare empirical human studies [198]. Furthermore, some works [172, 96] considered contrastive questions like "Why didn't you do ...?", but not contrastive explanations. In our thesis in the theoretical contribution, we generate contrastive explanations as a response to abnormal situations in the environment, *i.e.* the generation of contrastive explanations is context-aware. Additionally, contrastive explanations in their general sense are considered of both the beliefs (external events in the environment and internal agent states) and the intentions (goals the agent is committed to achieving) of the agent. Finally and on the practical side, the thesis conducts an empirical human study based on Agent-based Simulation (ABS).

3.3.1.3/ Combining Simplicity and Adequacy

Kulesza et al. [162] have extended their approach (*see* Section 3.3.1.1) that allowed human users to build mental models of the agents. In this extension, they have especially focused on how the soundness (*i.e.* nothing but the truth) and completeness (*i.e.* the whole truth) of the explanations impact the fidelity of humans' mental models of how a recommender agent works [162]. Their findings suggest that completeness is more important than soundness, *i.e.* increasing completeness via certain information types helped participants to generate better mental models. They also found that the oversimplification, as per many commercial agents, can be a problem, *i.e.* when soundness was very low, participants experienced more mental demand and lost trust in the explanations, thereby reducing the likelihood that participants will pay attention to such explanations in the first place.

Their model investigated the "sweet spot" between simplicity, *i.e.* simple explanations with little information, and informativeness, *i.e.* complete explanations with too much information. They have evaluated by training their recommender system and then performing a human study with only 17 participants. The result surprisingly showed that there is no sweet spot and that the solution is simply to give all the explanations possible to the hu-

man. In addition to the questionable validity of the results from a statistical point of view, one possible reason for such result is the chosen settings 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. Indeed, as confirmed in the literature, there is a need for harmonizing the explanations to the context and human information processing capabilities [214]. In summary, Kulesza et al. [162] have tackled a trade-off between soundness and completeness that could be viewed as an implicit combination of simplicity and adequacy. However, in our thesis, we explicitly make the combination of simplicity and adequacy (*see* Chapter 6).

In all cases, the related works showed the importance and need for performing empirical human studies to evaluate the proposed models and architectures with the help of humans that can subjectively determine if the explanations increase their understandability and trust. The next section discusses an interesting work as an example of such an empirical human study.

3.3.2/ Empirical Human Studies in Explainable Artificial Intelligence

Empirical human studies are vital to assess the process of explanation reception (*see* Section 2.7). Yet, very few works in the literature undertake such studies [198]. Examples of these works are [160] and [185]. The latter is a very recent and interesting work by Madumal et al. [185] that have investigated different levels of explanations (none, detailed, and abstract) for reinforcement learning agents. The work is based on the idea that prominent theories in cognitive science propose that humans understand and represent the knowledge of the world through causal relationships. In making sense of the world, humans build causal models in their minds to encode cause-effect relations of events in the environment and use these to explain why new events happen by referring to things that did not happen. In their work, the authors use causal models to derive causal explanations of the behavior of model-free reinforcement learning agents. Their approach learns a structural causal model during reinforcement learning and encodes causal relationships between variables of interest. This model is then used to generate explanations of behavior based on a counterfactual analysis of the causal model. They performed an empirical evaluation using a Human-Computer Interaction (HCI) study where 90 participants watch agents playing a real-time strategy game (Starcraft II) and then receive explanations of the agents' behavior. Later, the participants fill a questionnaire to collect their responses in terms of explanation quality and trust.

The results of the work of Madumal et al. [185] show that their model of abstract causal explanations offers 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 comparing their model of explanation with the same scenario but with no explanation, the results show no significance for the explanation quality metrics (complete, understand, and satisfying) and only manage to get significant results in the "sufficient details" metric. Additionally, in terms of the explanation trust metrics (confident, predictable, reliable, and safe), the obtained $p - values$ are not statistically significant using pair-wise Analysis of Variance (ANOVA) parametric test. Moreover, 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 understood the simulation, is significant for the model in this related work, while the 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. Furthermore, it is worth mentioning that the benchmark they have used in their work was mainly defined by them due to the lack of benchmarks in the domain. Finally, the filtering used in this related work was static filtering that does not change adaptively according to the context, unlike the proposal of the thesis. In our thesis, we extend the way of filtering into adaptive filtering (none, detailed, filtered) that is context-aware of the situation (*see* Section 6.5.2 on page 98).

3.3.3/ DISCUSSION

Despite considerable advances, the domain of XAI is still in its early stages of development, and particularly in the goal-driven XAI. Consequently, there are some open research issues and limitations to be tackled that we could categories as follows:

- **Agent Typology:** The organizing of the works in terms of typology can be based on different aspects. In terms of localization, the agents are categorized into remote agents and co-located ones. Another categorizing aspect could be between virtual agents and embodied ones: agents that interact with the environment through a physical body within that environment. These are some of the cases when discussing human-agent explainability. However, another type of explainability is agent-agent explainability that has its own challenges. Additionally, many works focused on few domains in their evaluations, *e.g.* simple games. The point of this category is that the challenges differ based on the different aspects and types, and the results attained in one aspect could not be easily generalized into the others.
- **Parsimony:** Several works tried to tackle concepts related to parsimony like the simplicity [122, 44, 279, 75, 187, 63, 186, 119, 141, 142, 45, 121, 120, 160] and the adequacy [145, 60, 202, 241, 302, 214, 56, 268, 172, 96]. However, there are still some open questions like how to achieve simplicity of explanations without

the risk of oversimplification? and how to achieve the adequacy of explanations without overwhelming the human with extra unnecessary details? In this direction, the idea of investigating a trade-off between simplicity and adequacy is promising, considering of course the risks associated with such a trade-off [162].

Most of the works tried to investigate the phases of providing an explanation (generation, communication, and reception) separately, *e.g.* investigating types of generation of explanations or ways of communicating the explanations. Therefore, there is a lack of works trying to build a combined solution that considers the concepts of parsimony in an explanation process, *i.e.* how to handle the phases of explanations in conjunction in one process that could benefit from the merits of each phase and compensates for its drawbacks.

Regarding the communication of explanations, most of the works tackled the textual visualization of explanation, while only some works considered multi-model approaches where several means of communication could be used, *e.g.* graphical, vocal, visual, simulated, *etc*. Even with the textual approaches, the issue of verbose explanations is not fully investigated. To tackle the latter issue, the filtering of explanations has been identified as a good way to increase simplicity when communicating the explanations to the human [120, 160, 141]. However, most of the related works in the literature considered static filtering of explanation that is based on simple predefined rules in the design time. In such case, it could be more beneficial to adopt a more adaptive way of filtering the explanations that could handle changes in the run-time, *e.g.* changes in the environment. Adaptive filtering could be also based on a user model: a model the agent builds about the human user. The latter adaptation considers the preferences of the human, *e.g.* providing explanations depending on the user age [141], or personalized recommendations based on the user preferences [235]. Abdulrahman et al. [3] have also started to tackle the issue of user-aware models in their very recent works [3, 2]. Even though some works started to work in this direction, the proposals and the evaluations are still in their first steps, and relatively little research has been conducted for personalized explanations [16].

- **Architectures and Models:** According to the SLR conducted by Anjomshoae et al. [16], the majority of the related works have not explicitly expressed their method for generating explanations. Additionally, several works relied on ad hoc methods to address their explanations problem. Following that, BDI architectures were implemented to generate explanations for goal-driven agents (*e.g.* [45, 214]). The less-used rest of the architectures and models include Markov Decision Process (MDP), Neural Networks (NN), Partially Observable Markov Decision Process (POMDP), and others [16]. Even though BDI has been the most used model, there is still a need to confirm its usefulness along with cognitive architectures for modeling ex-

plainability in several domains, *e.g.* remote robots. More importantly, architectures and models lacked the context-awareness when generating explanations. The definition of context-aware explanations is stated in Definition 7 on page 19.

- **Evaluation:** Some works have started to provide metrics to evaluate contributions of human-agent explainability [132]. However, the main limitation of the works in the domain of goal-driven XAI is the lack of empirical human studies to evaluate models and architectures addressing human-agent explainability [198, 16]. Moreover, very few works have conducted statistical analyses when analyzing the subjective outputs resulted from the performed human study. Furthermore, there is still a need for a standardized framework for evaluation and assessment that include test-beds well established in the domain.

This thesis handles some of the mentioned open issues as Research Questions (RQs). We choose an issue from each category keeping in mind that the chosen issues should be related and consistent with each other. To tackle the chosen research issues, we define a research methodology that outlines our approach of the theoretical contribution and the statistical evaluation of the contribution. The research methodology along with a more detailed discussion of the RQs (*see* Section 5.2) and the research hypothesis (*see* Section 5.3) handled in this thesis can be found in Chapter 5.

Investigating the literature leads us to discuss the parsimony of explanations (refer to Section 2.5 for more details) that could help in simplifying the explanations with different explanation communication techniques while keeping all the necessary information. This thesis argues that enforcing measures of parsimony of explanations helps to meet the two key features of simplicity and adequacy. This proposition will be further discussed in Chapter 6.

3.4/ Conclusion

A very recent SLR has been conducted in the domain of goal-driven XAI [16] showing an increased interest of researchers on this domain in the last five years (2015-2019). Two main findings were: *(i)* Several papers propose conceptual studies, or lack evaluations or consider relatively simple scenarios; *(ii)* while providing explanations to non-expert users has been outlined as a necessity, only a few works tackled the issue of context-awareness. The SLR found out that the most cited social science theory as a background is folk psychology. Moreover, the SLR stated that besides ad hoc solutions, the most used architecture to generate explanations for goal-driven XAI is the BDI architecture.

Recently, works in the XAI literature have started to respond to the two key features of an explanation (simplicity and adequacy). In particular, the filtering of explanations

has been suggested to achieve simplicity [120, 160, 141]. Yet, the solutions offered in these works were not flexible enough to consider complex situations, *i.e.* the models and the experiments were not adaptive to changes in the environment and rather defined particularly for a specific situation. Moreover, there was a lack of determining the best granularity level (either detailed or abstract) of the explanation to avoid overwhelming the humans in various situations. Furthermore, for some works (*e.g.* [185]), the results revealed no significance for most of the explanation quality metrics when comparing the proposed models of explanation in these works in a scenario with the same scenario but with no explanation.

To achieve adequacy, several works [145, 60, 202, 241, 302, 214, 56, 268] have investigated contrastive explanations, firstly pinpointed by Lipton [175], to provide explanations containing the necessary information needed by the human especially in abnormal situations [198]. Nevertheless, most of the works in the literature have been carried out at the conceptual level with rare empirical human studies [198].

One work by Kulesza et al. [162] has tried to combine the two features and discussed the "sweet spot" between simplicity and adequacy. After a human study with only 17 participants, the result surprisingly showed that there is no sweet spot. One possible reason for such a result is the chosen settings of the experiment, as there was no challenging situation that provides too many explanations to overwhelm the human user; *i.e.* the work did not tackle the human cognitive load. Additionally, the empirical human study conducted in this work included only 17 participants so the significance of the results is questionable.

Generally, the main drawback of the works in the domain of goal-driven XAI is the lack of empirical human studies to evaluate models and architectures addressing human-agent explainability [198, 185, 16]. Other drawbacks or challenges include the communication of the explanations, issues related to the core AI running the system, and context-awareness [16].

Despite considerable advances, the domain of XAI is still in its early stages of development, and there are some open research issues to be tackled. This section outlined and categorized these open issues. To handle these issues, we define a research methodology that outlines our approach of the theoretical contribution and the statistical evaluation of the contribution. The research methodology along with a more detailed discussion of the research questions (*see* Section 5.2) and the research hypotheses (*see* Section 5.3) handled in this thesis can be found in Chapter 5. The contribution that gives answers to these questions is discussed in Chapter 6.

As stated in the general introduction chapter, this thesis is a part of the academic project UrbanFly and one of the goals of this project is to propose novel models for simulating UAVs in urban environments and smart cities. In this context, UAVs represent the remote robots explaining their behavior and actions to the human. Working with remote robots is

a challenging task, especially in high-stakes and dynamic scenarios such as flying UAVs in urban environments. Therefore, ABS has been introduced as a promising tool to allow for the simulation of UAVs in such environments. We employ ABS in an application of UAVs in the evaluation part of the thesis (Part IV). Accordingly, the next chapter presents an SLR of ABS for UAVs.

4

Agent-based Simulation of Unmanned Aerial Vehicles

4.1/ Introduction

Recently, the civilian applications of Unmanned Aerial Vehicles (UAVs) are gaining more interest in several domains. Due to operational costs, safety concerns, and legal regulations, Agent-based Simulation (ABS) is commonly used to implement models and conduct tests. This has resulted in abundant research works addressing ABS in UAVs. This chapter[1] aims at providing a comprehensive overview of this domain by conducting a Systematic Literature Review (SLR) on relevant research works addressing ABS in civilian UAV applications in the previous ten years. This SLR aims to identify the most important questions and analyze the literature. To the best of our knowledge, no systematic literature study has been conducted to review the research addressing ABS in civil UAV applications.

The rest of this chapter is organized as follows: Section 4.2 states and defines the SLR methodology adapted from [46, 152]. Section 4.3 details the analysis and results of the SLR. Section 4.4 surveys the related works, and finally, Section 4.5 concludes this chapter identifying the key research perspectives related to the contribution of the thesis.

4.2/ Systematic Literature Review Methodology

Recently, research on computer science in general and on artificial intelligence, in particular, has witnessed a significant increase both qualitatively and quantitatively. For this reason, SLRs are becoming popular to help analyze the evolutions of these domains. Kitchenham and Charters [150] define SLR as follows:

[1] This chapter is based on our work [207].

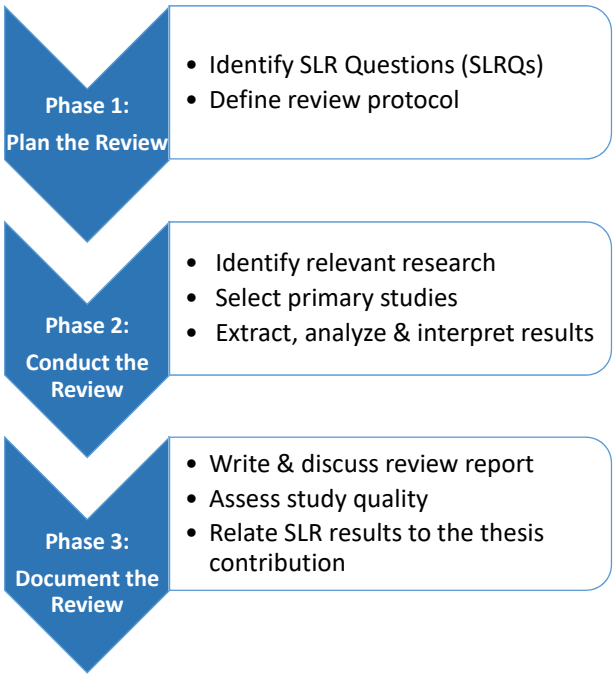


Figure 4.1: The systematic literature review process (adapted from [43, 102])

Definition 14: Systematic Literature Review (SLR) according to Kitchenham and Charters [150]

A form of secondary study that uses a well-defined methodology to identify, analyze and interpret all available evidence related to a specific research question in a way that is unbiased and (to a degree) repeatable.

Where secondary study refers to "a study that reviews all the primary studies relating to a specific research question." In this chapter, we define a primary study as a research work addressing a specific research question in the domain of UAVs. The aim of SLRs can be threefold [150]: *(i)* to summarize the existing evidence concerning a specific technology that is being used broadly, *(ii)* to identify gaps in the existing research to suggest areas for future investigation, and *(iii)* to provide a background allowing to position new research activities.

With these goals in mind, we base our SLR on [46, 152], which are among the most common methodologies for computer science SLRs. Such an approach ensures rigorousness, fairness, and reproducibility. Figure 4.1 illustrates the review process.

This section is organized as follows. First, Section 4.2.1 highlights the SLR questions. Second, Section 4.2.2 explains the review protocol, how conflicts are resolved and biases overcome. Third, in Section 4.2.3, the defined protocol is executed and the review process is undertaken (document collection, conflict resolution, *etc.*).

4.2.1/ SYSTEMATIC LITERATURE REVIEW QUESTIONS

Following the Goal Question Metric (GQM) [152], we define our generic free-form question as "Discover and evaluate the possible scientific Multi-agent Systems (MAS) contributions to the civilian UAV applications." On the field, it may be hard to deploy UAVs because of safety and security issues. Moreover, it may be difficult to reproduce the same scenario several times to test hypotheses and validate the behaviors of UAVs. ABS is a suitable tool for overcoming these limitations. to reflect this aspect, the generic free-form question becomes 'Discover and evaluate the possible scientific ABS contributions to the civilian UAV applications." This question is broken down into further Questions, that we abbreviate as SLR questions (SLRQs), exploring key issues in ABS for civilian UAV applications. The SLRQs are mainly concerned with this type of application. More specifically, these questions cover the purposes, issues, used simulation frameworks, publications date, authors, countries, *etc*. These questions were formulated based on the authors' knowledge in the UAV and ABS domains as well as the common practices from other SLRs. In what follows, 8 SLRQs are considered within this review, and we list all of them below:

SLRQ1 Identify the artificial intelligence models scaffolding the solutions in the reviewed papers. This SLRQ is set up to have a view on the types of agent architectures used for implementing civilian UAV applications. It also helps researchers understand the potential of these agent & system architectures and their limitations. The following sub-questions focus on specific types of models and architectures.

- SLRQ1-1 Investigate the agent architecture used in the solution (*e.g.* cognitive agent, Belief-Desire-Intention (BDI) agent, reactive agent, *etc*.).
- SLRQ1-2 Investigate the architecture of the system in the studied papers (decentralized vs centralized).
- SLRQ1-3 Investigate whether the proposed model (agent or simulation architectures) includes the environment, and how the UAVs interact with their environment.

SLRQ2 Identify the main model category used by the proposed work (*e.g.* mathematical, algorithm-based, *etc*.). This SLRQ is set up to determine if the contributions to civilian UAV applications are formal or semi-formal. It will lead us to a statement and arguments regarding the validation of UAV models.

SLRQ3 Identify the simulation frameworks used to implement the proposed solutions, and the main advantages & disadvantages of each framework especially if it excels in a specific civilian UAV application domain. This SLRQ is related to the technological means used by the researchers for implementing the MAS for civilian UAV applica-

tions. It should help researchers to determine and choose the best framework for their model implementation.

SLRQ4 Understand the evolution of UAV simulations in MAS in the last decade in terms of key contributors (research labs), geographic distributions, growth over the years. Having answers to this SLRQ will help researchers determine the liveliness of the MAS modeling domain for civilian UAV applications. Moreover, this SLRQ will enable highlighting the active contributors. In this way, it may help researchers to find quickly new contributions in the domain.

SLRQ5 Identify the main UAV research topics and civilian application domains addressed in the studied papers. On one hand, we consider it is important to determine active research topics to highlight less active research topics where more contributions are needed. On the other hand, and as the market of civilian UAVs is expanding rapidly in several application domains [280], a synthetic view of these application domains will enable researchers to determine the typical applications for their research, and possibly identify new application domains.

SLRQ6 Investigate whether models and technologies enabling to implement concepts related to Internet-of-Things (IoT), pervasive systems, or ubiquitous systems are considered in the studied papers. IoT is a technological domain that is more and more used within smart cities. This SLRQ contemplates if technologies like wireless sensor networks, connected vehicles, connected buildings, etc. are considered as a component of the agent-based system in the studied papers. In such systems, IoT contributes to the model at the agent level (objects may be modeled as agents) and at the agent environment level (objects are not agents). This SLRQ will also highlight how IoT devices and UAVs are interacting together.

SLRQ7 Identify the communication technology used by the UAVs to connect to other entities: For UAVs to be deployed and used in their environment, especially in smart cities, they could either be connected to infrastructure entities, *i.e.* Vehicle-to-Infrastructure (V2I), or to other UAVs, *i.e.* Vehicle-to-Vehicle (V2V). Understanding the communication technology of the UAV in the proposed simulated works is key to assess whether frameworks are capable of producing realistic simulations.

SLRQ8 Assess the evaluation of the proposed model: Simulation for civilian UAV applications needs scenarios to be set up, and the use of datasets may help to create such simulation realistically. This SLRQ assesses if the evaluation relies on a dataset, on a generated synthetic dataset, or no dataset. This will highlight the different datasets and the scenarios (if no dataset is used) from the literature.

In this chapter, we answer only 3 of the SLRQs that are the most related to the contribution

of the thesis, namely SLRQ1, SLRQ2, and SLRQ3 that are directly related to the thesis. The rest of the SLRQs answers are detailed in Appendix A.

4.2.2/ Defining the Review Protocol

As shown in Figure 4.1, defining the review protocol is done right after having set the SLRQs. The protocol we used for this SLR involves the following steps. First, Section 4.2.2.1 chooses the databases used as sources of information and defines the stop criterion. Section 4.2.2.2 defines the exclusion/inclusion criteria used by the reviewers to exclude/include articles chosen from the databases before the stop criterion was triggered. Section 4.2.2.3 presents the quality criteria used by the reviewers to assess the quality of the primary studies. Finally, Section 4.2.2.4 explains the policies used to mitigate subjective biases and resolve conflicts.

4.2.2.1/ Database Selection

This process is composed of the following couple of steps:

1. IEEE Xplore, ACM Digital Library, and Google Scholar are selected as the three databases constituting the source of information. The selection of the first two databases is obvious in computer science. Google Scholar is selected because it provides a large list of documents that are not indexed into the two previous databases, *e.g.* papers from conference proceedings, Ph.D. and Master theses. Despite not being peer-reviewed, these articles obtained from Google scholar might be important given that the interest in the studied topic is rapidly increasing in recent years.

2. The databases are queried with a set of keywords. These keywords are devised based on the authors' knowledge of the UAV and ABS domains.

When queried with these keywords, each database responded with a set of articles that are considered by the reviewing process. The number of articles to be produced by the queries is relatively large for IEEE Xplore, ACM DL, and Google Scholar databases. However, only a few of these articles were relevant to the SLRQs raised in the previous section. For this reason, as in [50], the following stop criterion was applied: "Stop the collecting of articles after a sequence of 10 titles, completely incoherent with the query, appeared in the list." Determining whether an article is coherent is left to the reviewers' subjective view when they deemed that there was no adherence between the query performed on the database and the title/abstract of the article appearing in the result.

4.2.2.2/ Inclusion and Exclusion Criteria

The papers appearing in the resulting pool of papers are not necessarily useful to answer the SLRQs defined above. For this reason, most of the literature review methodologies [102, 50] apply a set of exclusion criteria to retain only pertinent papers. The set of exclusion criteria, defined by the authors, is listed below:

ExC1 Not a recent research work — Papers that were published before 2008, *i.e.* with a publication year < 2008, are excluded. It is assumed that the non-recent research is not up-to-date due to the high evolution rate of UAV technologies and usages.

ExC2 Invalid type of paper, the document is a poster or a demo — It is assumed that a poster or a demo cannot give enough details on the contributions, as there is no enough contributed content for evaluation. Ph.D. theses, Master theses, technical reports are included.

ExC3 Invalid type of paper, the paper is a survey — It is assumed that the survey papers (*i.e.* secondary studies) do not provide contributions directly on the UAV models nor UAV technologies.

ExC4 Impossible to access the paper text — It is impossible to evaluate a paper when its text cannot be accessed (PDF download, online text, *etc.*).

ExC5 Extended paper — The paper is extended by another paper by the same authors. The contributions in the extended paper are enclosing the ones from the original paper so that the latter is excluded.

ExC6 Unrelated to UAV — The paper has neither a contribution in the fields of UAV models nor UAV technologies.

ExC7 Unrelated to agent-based systems – The paper has neither a contribution in the fields of agent-based technologies nor distributed artificial intelligence. Generally, only multi-agent applications are included, but if the system includes agents that communicate with other entities like Infrastructure (refer to ExC14), then they will also be included.

ExC8 UAV manufacturing only — The paper's contribution is related to the manufacturing of UAVs, *i.e.* it is related to the design and implementation of hardware, mechanic, or electronic components.

ExC9 Positioning system only — The paper's contribution is related to the definition of novel positioning systems within UAVs. The contribution focuses on a perception model that enables each UAV to compute its position in the air.

ExC10 UAV detection system only — The paper's main concern is UAV detection within the system. In other words, the contribution is not related to UAV behavior, but to a system that is detecting the UAV in the air.

ExC11 No civilian application — The paper contains only military applications that cannot be applied to civilian fields.

ExC12 No simulation contribution — The paper's contribution cannot be applied to UAV simulation. In several papers, the model is deployed on real UAVs without simulation. Even if the paper has not a direct contribution to UAV simulation, if the proposed model could be deployed within a simulation environment, the corresponding paper is not excluded.

ExC13 Simulation in 2D — The paper's contributions include a simulation model in 2D that cannot be extended to the third dimension. It is assumed that 2D simulation of UAVs that cannot be extended into 3D cannot achieve the highly detailed reproduction of the UAV behavior when they are in the air. However, to estimate the portion of 2D and 3D simulations, we keep track of this exclusion criterion (*cf.* Section 4.2.3).

ExC14 No UAV cooperation nor interaction — The paper contains a contribution related to neither the cooperation of UAVs nor the interaction between UAVs. V2I and V2V communications are assumed to be the base framework for supporting UAV interaction. If a paper contains a model for a single UAV that has V2I communication, it is not excluded since this type of model could be duplicated to set up a more complex simulation environment based on stigmergy[2] communication.

ExC15 No autonomous UAV — The paper contains only a contribution related to the piloting or controlling the UAVs. In other words, the contribution does not focus on the autonomous behavior of the UAVs.

These exclusion criteria are applied to the documents in two steps. In the first coarse-grained step, the articles were only eliminated if their titles and abstracts satisfied at least one of the exclusion criteria. In the second fine-grained step, the remaining papers are screened but this time reading the whole body of the paper.

4.2.2.3/ QUALITY CRITERIA

As has been recommended by Kitchenham and Charters [150] and Kitchenham et al. [151], most of the SLRs rely on quality criteria allowing to assess the quality of primary

[2]This term has been introduced by the French biologist Pierre-Paul Grassé in 1959 for describing the termite behavior. It is defined as: "Stimulation of the workers by the work that they performed." This term expresses the notion that the actions of an agent leave signs in the environment. These signs are perceived by itself and other agents, which determine their next actions [224].

studies (*e.g.* [102, 50]). Defining quality criteria as a list of questions is a common practice. Typical quality criteria include: *(i)* whether the authors of primary studies provided a sound rationale for their work, *(ii)* details about the context and the design of the technical evaluation, *(iii)* the statement of the results.

Note that, as it is the case in [102], the quality criteria are not used to exclude/include primary studies. Rather, they are used to report the overall quality of primary studies included by the SLR. To assess the quality of the reviewed works, Table 4.1 defines four quality questions, adapted from [102]. Note that Q3 is of particular interest since having an overview of the quality of evaluations of the set of articles dealing with a specific research question can give a good idea on the maturity of this research question.

#	Quality Question
Q1	Do the authors provide a sound rationale (*i.e.* motivation) for their work?
Q2	Is there an adequate description of the context in which the study has been conducted?
Q3	Is there a clear statement of the findings and the results including data that support the findings?
Q4	Are the limitation of the study discussed and highlighted?

Table 4.1: The Quality Questions

4.2.2.4/ Biases and Disagreements

to mitigate the subjectivity of the reviewing process, certain measures were taken to overcome biases and resolve conflicts. In particular, each task of Phase 2 in Figure 4.1 was conducted by at least 2 reviewers. Thus, as shall be discussed later, the steps of article exclusion/inclusion (*see* Section 4.2.2.2), answering the SLRQs, and quality assessment (*see* Section 4.2.2.3), were undertaken by at least two reviewers for each article. A third reviewer intervened as a referee to resolve a conflict in the exclusion/inclusion and the SLRQ answering steps. As for the quality assessment task, quality assessments given by reviewers for each article were averaged.

4.2.3/ Performing the Review

This section gives an account of how the SLR has conducted and analyzed the results of the exclusion/inclusion step. Two keyword sets were applied to the three databases (IEEE Xplore, ACM DL, and Google Scholar). The first keyword set is {`UAV`, `agent`, `simulation`}, and the second set is {`UAV`, `agent-based simulation`, `drones`, `civil`, `multi-agent systems`}. A stop criterion of 10 articles was applied to the results of each keyword set and database. After the stop criterion was applied, the total number of

articles retained was 316. The next step is to apply the coarse-grained exclusion/inclusion step. Note that since this step screens papers based on their titles and abstracts, some exclusion criteria might be more helpful than others (*e.g.* ExC1 and ExC2).

Database	Number of Papers	Percentage
Selected from Keyword Set 1	131	≈ 41%
Selected from Keyword Set 2	185	≈ 59%
Sum from Set 1 & 2	316	=100%
Total Included	123	≈ 39%
Total Excluded	193	≈ 61%

Table 4.2: The results of the coarse-grained exclusion/inclusion step

	Number of Papers	Percentage
Included	30	≈ 24%
Excluded	70	≈ 57%
Conflict	23	≈ 19%
Referee Included	12	≈ 10%
Total Included	42	≈ 34%
Total Excluded	81	≈ 66%

Table 4.3: The results of the fine-grained exclusion/inclusion step

Table 4.2 shows the results of the coarse-grained exclusion/inclusion step. As can be seen from the third row in the table, the total number of papers acquired from the two sets combined is 316 papers. The results listed in Table 4.2 reveal that about 39% (123 papers) of the total number of papers were included by the coarse-grained exclusion/inclusion step.

The next step within the review process is the fine-grained exclusion/inclusion step. It is applied to the 123 papers selected during the previous step. More specifically, the content of the paper is screened and the paper is excluded if it satisfies at least one of the exclusion criteria defined in Section 4.2.1. Table 4.3 shows the results of the fine-grained exclusion/inclusion step. Each paper was reviewed by at least two reviewers. If all the reviewers of a paper decided that it should be included in the review, the paper is included (*cf.* the first row of the table). The paper is excluded if all of its reviewers agreed upon its exclusion (the second row of the table). Otherwise, in case of a conflict among reviewers, we relied on a referee to resolve this conflict. If the referee accepts the paper, then it is included in the review. As shown in the table, 23 papers (about 19% of the total number of the papers) generated conflicts among the reviewers. Out of these papers, 12 papers were added by the referee raising the total number of included papers to 42 (about 34% of the papers remaining after the coarse-grained exclusion/inclusion step). For a list of the 42 papers, please refer to Table 4.4.

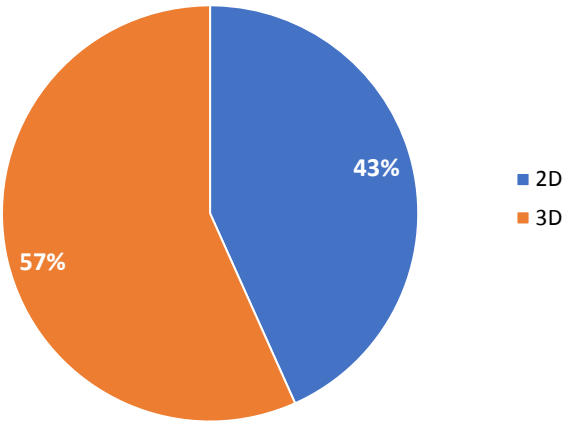


Figure 4.2: The percentage of papers with a 2D simulation scenario vs. papers with a 3D simulation scenario

Most of the papers screened in the fine-grained step were highly related to the SLRQs. As indicated by Table 4.3, 66% of them were excluded. Note that some papers were excluded because of satisfying multiple exclusion criteria.

Figure 4.2 compares the percentage of papers presenting 2D and 3D simulations in the pool of papers before the fine-grained exclusion/inclusion (whose total number is 123), excluding 26 papers that were not determined either they use simulation or not, or either they are 2D or 3D. to understand the general tendencies of 2D and 3D simulations, Figure 4.3 plots the number of papers proposing 2D/3D simulations per year. As shown in the figure, the number of papers proposing 3D simulations is witnessing a confirmed and a significant increase (2018 should not be considered since this review was conducted in August 2018).

This section offered a detailed account of how the review was performed and provided useful statistics about the included/excluded papers. Furthermore, it discussed the most common exclusion criteria. the next section presents and analyzes the results of the SLR regarding the SLRQs presented in Section 4.2.1.

4.3/ Results and Analysis of the Review

This section thoroughly analyzes the SLR results. It analyzes the papers retained after the fine-grained exclusion/inclusion step and discusses 3 out of the 8 SLRQs defined in Section 4.2.1, namely SLRQ1, SLRQ2, and SLRQ3 (page 43). These 3 SLRQs have been chosen because they are directly related to the contribution of the thesis[3]. Note that the results are related to ABS in civilian UAV applications and are derived from the SLRQs and the exclusion criteria defined in the SLR methodology.

[3] *see* Appendix A for SLRQ4, SLRQ5, SLRQ6, SLRQ7, and SLRQ8.

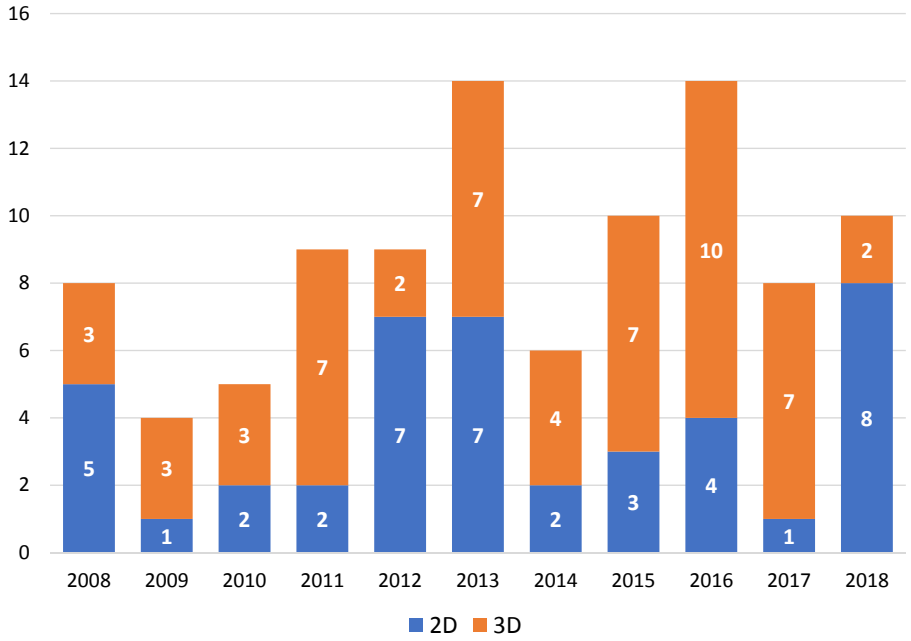


Figure 4.3: The number of papers with 2D and 3D simulation scenarios per year

4.3.1/ Artificial Intelligence or Agent Architecture Type (SLRQ1)

This section deals with the SLRQ1. First, in Section 4.3.1.1 the used agent architectures are identified and discussed. Second, in Section 4.3.1.2, the used system architectures are analyzed. Finally, Section 4.3.1.3 explores how the primary studies dealt with the dynamicity of the environment.

4.3.1.1/ Agent Architecture (SLRQ1-1)

The results of the SLR concerning the agents' architectures showed that agents used in the studied research works, mainly fall into five categories: *(i)* Reactive agents, *(ii)* Flocking agents, *(iii)* Belief–Desire–Intention (BDI) agents, *(iv)* Agents using cognitive architectures, and *(v)* Evolutionary agents. The following list presents these architectures respectively.

- **Reactive agents:** The behavior of reactive agents is driven by their reactions to external stimuli (*e.g.* a message from another agent) or a change in their environment (a perceived obstacle).

  Within the reactive agent behaviors, **Flocking agents** behavior is the behavior exhibited when a group of agents, *e.g.* birds, fishes are moving together. Basic architectures of flocking behavior are controlled by two simple rules: *(i)* alignment for steering towards the average heading of neighbors, *(ii)* cohesion for steering towards the average position of neighbors. With these two simple rules, the flock

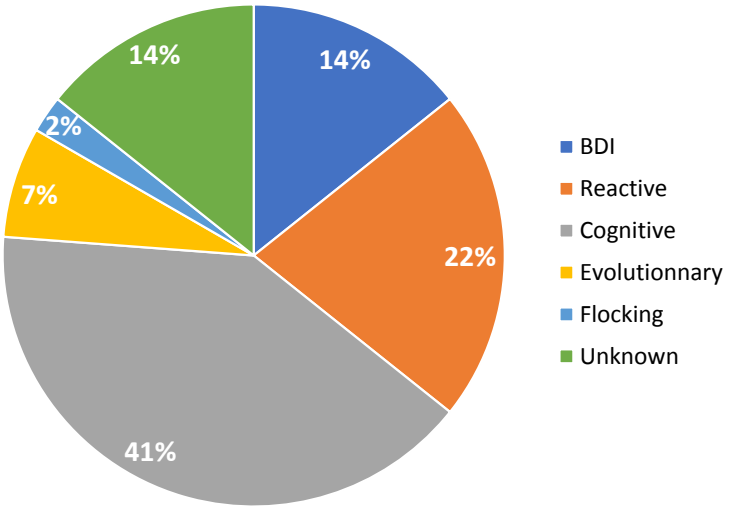


Figure 4.4: The agent architectures used in the reviewed papers

moves in an extremely realistic way, creating complex motion and interaction that would be extremely hard to create otherwise.

- **Cognitive agents:** Agents of this type rely on cognitive architecture. The latter aims at describing human cognitive processes as precisely as possible. In contrast to BDI, whose main inspiration is philosophical and relies on Michael Bratman's theory of human practical reasoning and on modal logic [41], other cognitive architectures are inspired by an in-depth understanding of the human brain from biological and neurological perspectives. There are many implementations of cognitive architectures (*see* 2.8). Soar [165] is one of the widely used ones.

  A special type of cognitive agents is **BDI agents**. They are rational agents that having a "mental attitudes" of Beliefs, Desires, and Intentions representing respectively the information, the motivational, and the deliberative states of the agent [238]. BDI agents are capable of integrating planning, scheduling, execution, information gathering, and coordination with other agents [274].

- **Evolutionary agents:** They are agents that are based on evolutionary algorithms. An evolutionary algorithm is a subset of evolutionary computation [296], a generic population-based metaheuristic optimization algorithm. It uses mechanisms inspired by biological evolution, such as reproduction, mutation, recombination, and selection. Candidate solutions to the optimization problem play the role of individuals in a population, and the fitness function determines the quality of the solutions. The evolution of the population then takes place after the repeated application of the above operators.

Figure 4.4 depicts the agent architectures used in the papers. As can be seen from the figure, cognitive architectures were the most common among the primary studies ($\approx 41\%$)

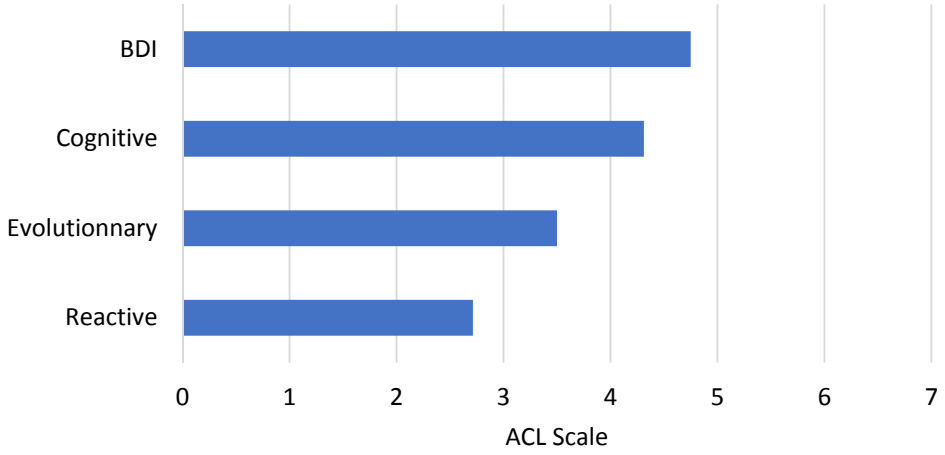


Figure 4.5: The average ACL per agent architecture

followed by Reactive agents ($\approx 22\%$) and BDI ($\approx 14\%$). This result reveals that proactive agents, those agents capable of goal-driven behavior (BDI agents, and cognitive agents) constitute about 55% of all the analyzed studies whereas reactive agents (including those with flocking behavior) are less common ($\approx 24\%$ of the analyzed works). This shows that most of the research works seek to equip the UAVs with greater autonomy and goal-driven behavior.

to understand the correlation between the agent architectures and the autonomy, Figure 4.5 depicts the average ACL per agent architecture (*cf.* [66] and Section A.2). The BDI agents offered the highest level of autonomy, followed by cognitive agents, evolutionary agents, and lastly reactive agents. This result confirms our expectations and suggests that cognitive architecture and BDI agents are promising paradigms allowing to build more autonomous UAVs.

4.3.1.2/ DECENTRALIZATION/CENTRALIZATION (SLRQ1-2)

Figure 4.6 shows the number of papers with reactive/direct collaboration system model, and with the system architecture used by the model (decentralized/centralized). As expected, the majority of the papers are related to decentralized architectures, which correspond to one of the major characteristics of MAS and UAV systems. However, 12 papers contain proposals that correspond to a centralized architecture, *i.e.* the model contains a central agent or the model is formalized in such a way that it could be implemented only with centralized frameworks, *e.g.* Simulink [223] (*cf.* Section SLRQ3).

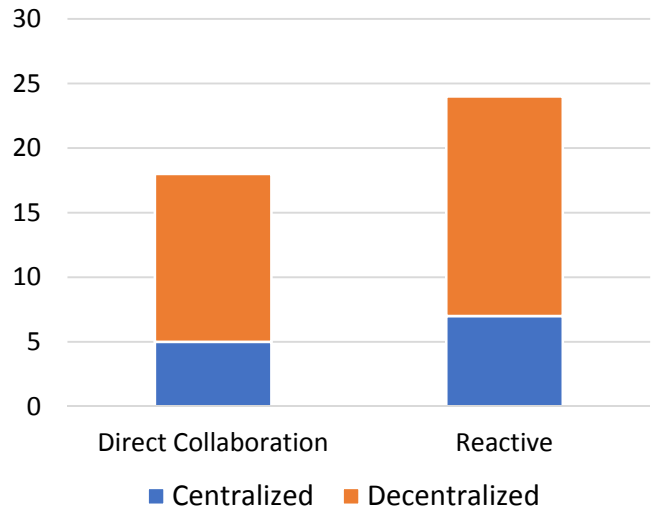


Figure 4.6: The number of papers with the system model (direct collaboration/reactive) and the system architecture (centralized/decentralized)

4.3.1.3/ Environment Dynamicity (SLRQ1-3)

Immersing agents in dynamic physical, virtual, or mixed environments is still a challenge for MAS researchers. As has been established in [311], an essential part of such systems is the MAS environment, to offer the services allowing agents to interact with it. However, defining what is the interface between the agents and their environment is not obvious. A key aspect is to respect their autonomy and ensure that the rules of the environment are enforced. Weyns et al. [311] define the agent environment as the software layer between the external world and the agents.

Dynamic agent environments include endogenous processes that enable the environment's state to evolve dynamically outside the control of the agent. In a static agent environment, such a process is not included. Additionally, the agent environment state could evolve only as a consequence of the agents' actions. If an action is never applied to the agent environment by the agents, it is passive.

Figure 4.7 depicts the proportion of dynamic and static agent environments. Half of the papers propose models based on a static environment, 48% in a dynamic environment. Due to the complexity of the UAV systems, static agent environments are used to control the complexity of the modeling and enabling an easier validation.

Agent environments in most of the reviewed papers are passive. In these cases, UAV missions are mainly surveillance, collision detection, coordination, *etc*. However, as UAVs become involved in more application domains where acting on the environment is necessary, there are likely to become more active in the agent environment.

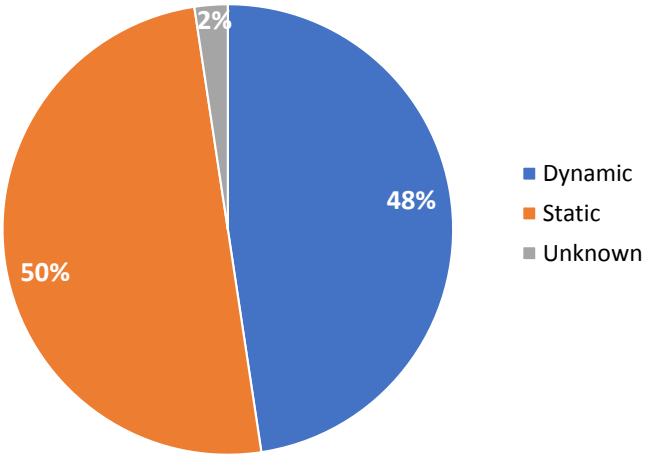


Figure 4.7: The proportion of dynamic and static agent environments

4.3.2/ FAMILY OF MODELS (SLRQ2)

The purpose of this SLRQ is to identify the family of the models that are used within the proposed works. Three major families are highlighted:

- **Mathematical model:** includes formal models that enable to verify and validate the behavior of the UAVs. Formal verification is the act of proving or disproving the correctness of UAV algorithms based on a certain formal specification, using formal methods of mathematics. The verification of these systems is done by providing a formal proof on an abstract mathematical model of the system, the correspondence between the mathematical model and the nature of the system being otherwise known by construction.
- **Algorithm-based model:** are the models based on the general computer programming theory. These models are instances of a logic written in a software to produce the behaviors of UAVs. These algorithms are not based on mathematical models, such that it is hard to give a proof of completeness and stability.
- **Not-categorized model:** If a paper's contribution can be classified neither mathematically nor as algorithm-based, it is put into the "None" category. In most of the reviewed papers, the contributions within this category are presented with abstract or general explanations without equations, algorithms, state machines, *etc*. For example, the UAV behavior is described by a schematic drawing.

Figure 4.8 illustrates a re-partition of the models according to their family. It is interesting to note that 38% of the proposed models are mathematical, and 43% are algorithm-based. Indeed, even if mathematical models are harder to define than algorithm-based models, safety concerns related to UAVs lead researchers to give a proof of safety and stability of the UAV behaviors over the time by providing mathematical models. Safety validation

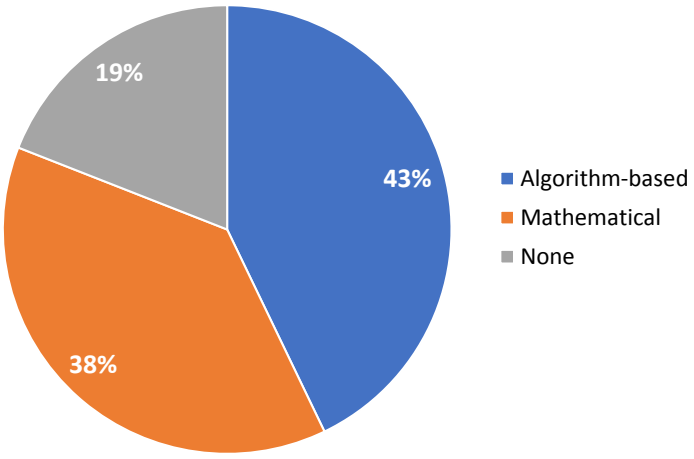


Figure 4.8: The main families of models

is not outside the algorithm-based models. In all the related papers, simulation testbeds are used for validating the behaviors of the proposed models.

As mentioned within the SLRQ1 analysis, the architecture in the reviewed models could be classified as centralized or decentralized architectures. Figure 4.9 shows the correlation between the system architecture classification and the mathematical/algorithm-based model classification. As expected for MAS, the architectures are mostly decentralized, whatever the type of model.

4.3.3/ Simulation Frameworks (SLRQ3)

Enabling early validation of a UAV system design requires the simulation of its components. This requires the development of an adapted simulation environment. Identifying the frameworks used to implement the proposed solutions, and the main advantages & disadvantages of each of them is a challenge by itself.

Figure 4.10 depicts the used simulation frameworks in the reviewed papers. Several frameworks are used: AgentFly [264], Simulink [223], Gazebo [153], NetLogo [313], MASON [180], A-globe [262], Repast Simphony [68], JADE [30], PROMELA [134], Gwendolen [76], Neptus [78], jME3 [294], and SPADE [111]. Table 4.4 provides a list of these frameworks. All used frameworks are open source except for AgentFly, A-globe, and Simulink. In our previous work [205], a comparison between open-source ABS frameworks was provided. However, the comparison lacked some frameworks that were revealed by the SLR presented in this work[4].

The two most used frameworks by the papers in the SLR are AgentFly (7%), Simulink (7%). However, the result indicates that no framework was favored by researchers for

[4] *see* Appendix B for the latest version of the ABS frameworks comparison.

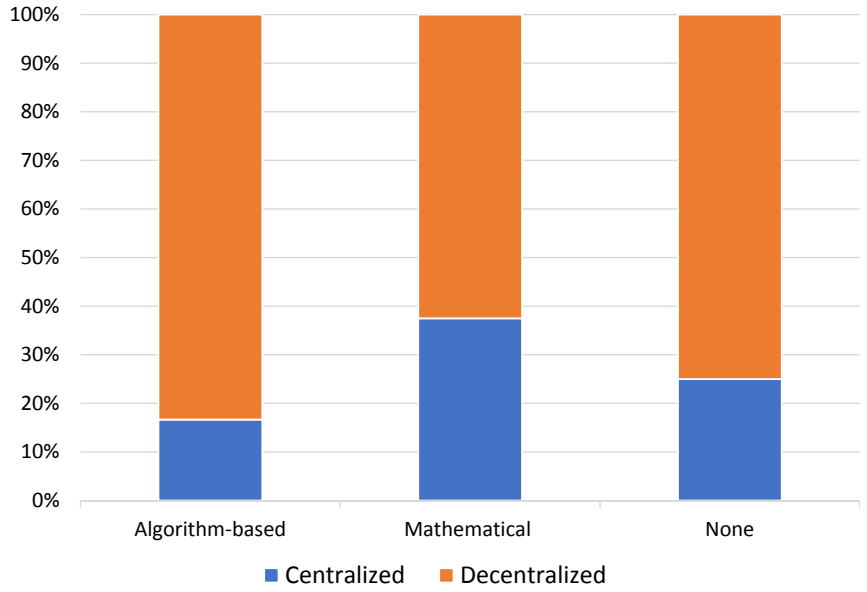


Figure 4.9: The main formal models correlated with the agent architecture (decentralized/centralized)

civilian UAV applications. The larger part of the implementations in the reviewed papers is using an ad hoc framework (43%). These simulation frameworks are typically developed by the authors of the reviewed papers from scratch. This fact leads us to consider that the existing frameworks do not cover all the needs mandatory for implementing a UAV simulation software. Moreover, and due to the abundance of frameworks with no clear distinguishing features related to UAVs, the authors generally prefer to set up their configuration and build the simulation framework from scratch even though it is time-consuming. Regarding the public availability of the implemented solution, only one paper has offered full access to the code [326], and some papers provided partial access to pieces of code or to open source tools that they used [248, 143, 292, 181, 294, 64, 249, 219], while the rest of papers did not offer any access.

4.3.4/ DISCUSSION

The results and the discussion above help to understand the recent tendencies in the studied domain. Nevertheless, the conclusions drawn in this chapter are only valid within the predefined domain of ABS and MAS for civilian UAV applications. Thus, the tendencies discussed in Section 4.3 cannot be generated to all UAV applications. Their scope and validity are limited by the keywords and the exclusion criteria defined in Section 4.2.

Based on the quality criteria defined in Section 4.2.2.3, the quality of the papers was evaluated according to the four quality criteria defined in Table 4.1 related to the explanations on the motivations (Q1), the study context (Q2), the theoretical and experimental results

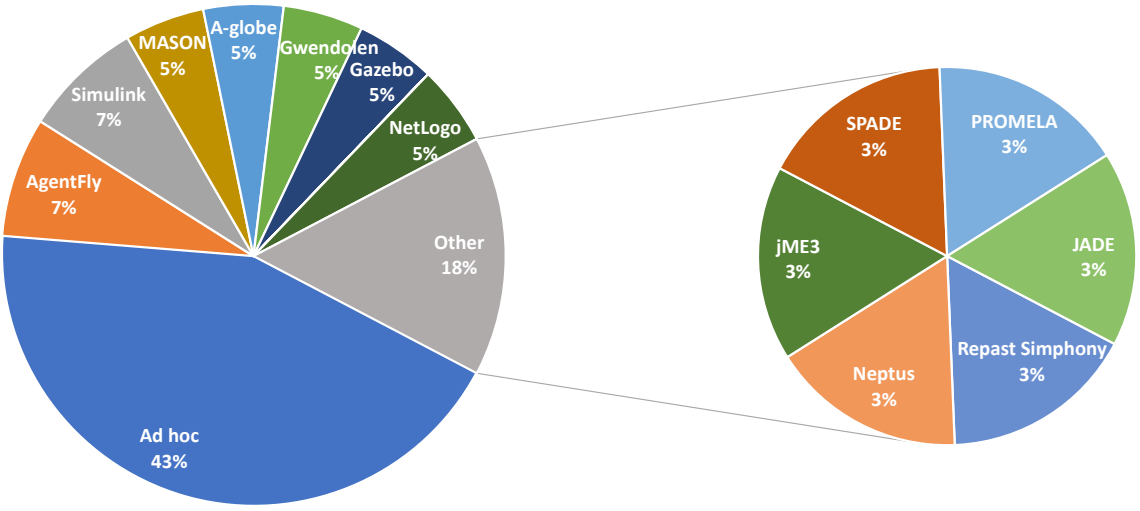


Figure 4.10: The used simulation frameworks

Framework	Papers using the framework
Ad hoc	Agogino et al. [7], Ashraf et al. [18], Wei et al. [307, 306], Bürkle et al. [49], Rollo et al. [243], Gunetti et al. [115], Evertsz et al. [86], Kandil et al. [140], Sampedro et al. [248], Peng et al. [230], Benedetti et al. [31], De Benedetti et al. [73, 74], Bürkle and Leuchter [48], da Silva et al. [261], Schatten [249]
MASON [180]	Albani et al. [9], Zou et al. [326]
AgentFly [264]	Semsch et al. [254], Pechoucek et al. [228], Šišlák et al. [263]
PROMELA [134]	Webster et al. [305]
A-globe [262]	Volf et al. [299], Stenger et al. [266]
NetLogo [313]	Cimino et al. [65], Zhu et al. [325]
Simulink [223]	Gunetti et al. [114], Ciarletta et al. [64], Kucherov and Kucherov [159]
Gwendolen [76]	Webster et al. [304]
Gazebo [153]	Arokiasami et al. [17], Ma et al. [181]
Repast Simphony [68]	Khaleghi et al. [143]
Neptus [78]	Vasilijevic et al. [292]
jME3 [294]	Veloso et al. [294]
SPADE [111]	Obdržálek [219]
JADE [30]	Fulford et al. [98]
n/a	Van der Walle et al. [301], Sutton et al. [271], Bentz and Panagou [32], Ferrag et al. [90]

Table 4.4: The reviewed papers per used framework

(Q3), and the limitations and research directions (Q4). Figure 4.11 shows the average evaluation for each of these criteria. Reviewers have provided a score, according to their background and knowledge, based on three levels of quality: “bad”, “average”, “good”. Each paper was evaluated by at least two reviewers and the results were averaged into

a scale from 1 to 5. Over the entire set of papers, motivations and contexts are clearly explained. The presentation of the results is described with the minimum set of details to allow a researcher to reproduce the presented results. Finally, as can be seen from the figure, reviewers have considered that the limitations of the proposed models and approaches are not enough detailed within the papers. However, as UAV technology evolves and new UAV manufacturers claim new customers and social flight clubs enlist fresh enthusiast amateur pilots, these open issues and limitations are becoming significant challenges.

There is a confirmed tendency towards the development of increasingly autonomous UAV systems. This evolution would minimize the human intervention by relieving him/her from the burden of continuously monitoring the UAVs. Nevertheless, in unpredictable situations, the UAV behavior might not conform to the expectations of the human operator. For instance, in a product delivery scenario, an autonomous UAV may choose to deviate from its expected path because of an unforeseen event. Enhancing the UAVs with explaining capabilities would allow the human operator to understand the reasons behind UAV behavior and raises its trust in the autonomous UAV system. Moreover, the recent developments of the domain of Explainable Artificial Intelligence (XAI) help UAVs to move in this direction. UAV has been cited as one of the applications where XAI would be needed [116]. Furthermore, developing explainable UAVs would have a very positive impact on human-machine teaming. Recently, this research direction is being explored by military UAV applications. Similar efforts should be considered for civilian applications [293]. As shown by recent studies, using BDI agents is a promising approach to develop explainable agents [45]. A key explanation for this success lies in the fact that the BDI paradigm is inspired by folk psychology (*see* Section 2.8). Therefore, BDI architecture offers a more straightforward description making models easier to explain for end-users. For this reason, UAV agents relying on BDI architectures, which represent only 14% of papers reviewed in this SLR (*cf.* Figure 4.4), are likely to increase in numbers if the issue of explainable UAV behavior is to become a hot research topic. This discussion confirms the research needs related to XAI that are detailed in Chapter 3.

4.4/ RELATED SURVEYS

Recently, several works surveyed the emerging topics of UAVs. However, these works mainly focused on vertical applications without considering the aspects and challenges across multiple application domains and research topics. For instance, Hayat et al. [125] focused on the characteristics and requirements of UAV networks for envisioned civilian applications between 2000 and 2015 from a communications and networking point of view. Motlagh et al. [204] reviewed Low-altitude UAVs highlighting their potential use

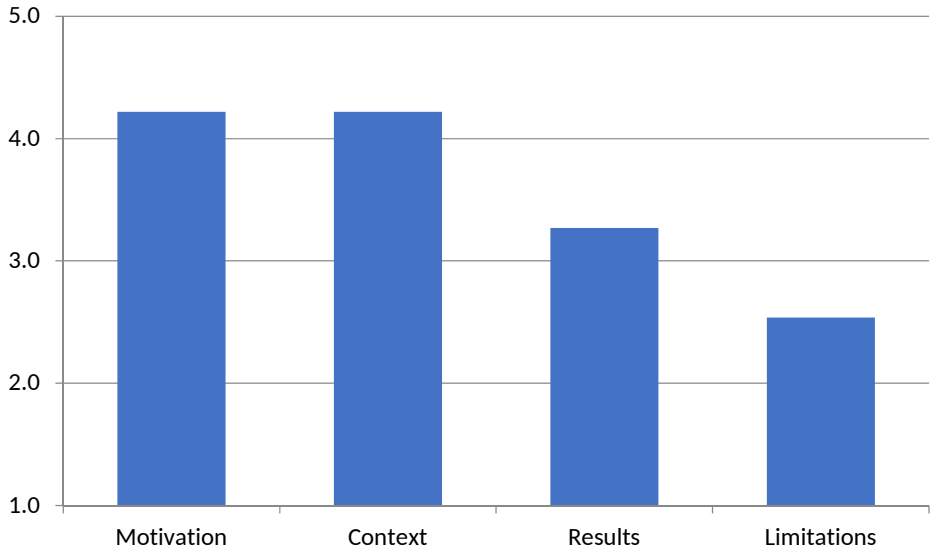


Figure 4.11: The qualitative evaluation of the reviewed papers

in the delivery of IoT services from the sky. Other surveys focused on traffic management [234], environmental monitoring [26], ad hoc networks in UAV applications [323], routing and energy efficiency in UAV communication networks [117], and UAV coverage [59].

One interesting survey is provided by Chmaj and Selvaraj [61] in which the authors surveyed the applications implemented using cooperative swarms of UAVs that operate as a distributed processing system. However, this survey did not tackle the challenges facing UAVs in these applications and the potential role of new technologies in UAV uses.

Shakhatreh et al. [258] reviewed civilian UAV applications and challenges. They identify current research trends and future challenges for civilian UAV applications, including: charging, collision avoidance, swarming, networking, and security-related challenges. Yet, this survey was mainly inspired by low-level aspects of UAVs like networking and wireless communication. Moreover, the listing of the comparison was in sequence without a cross-application domain discussion.

Other surveys focused on system identification and UAV-human interactions. In particular, current methods and applications of system identification for small low-cost UAVs were provided by Hoffer et al. [131], while the interaction between UAVs and humans applications was considered in another survey [155]. In this later work, a taxonomy of control methods that enable operators to control swarms effectively was developed. With highlighting challenges, unanswered questions, and open problems for Human-Swarm interaction.

4.5/ CONCLUSION

UAVs are becoming increasingly popular for civilian applications. The aim of this chapter is to conduct an SLR on research addressing MAS, and specifically ABS, in civilian UAV applications. This concluding section states the principal findings. Note that while the SLR concentrated on research using ABS for civilian UAV applications, some of the findings below pertain to key research issues in agent and MAS (*e.g.* agent architectures, decentralization, *etc*.).

Following a well-established SLR methodology, we have identified 8 SLRQs helping to assess the contributions of MAS and ABS in civilian UAV applications. The main findings of the most related 3 SLRQs (SLRQ1, SLRQ2, and SLRQ3) to the research questions (Chapter 5) of this thesis are[5]:

1. The majority of papers covered by the SLR opted for proactive agents, those agents capable of goal-driven behavior (BDI agents and cognitive agents) in their proposed model. Furthermore, most of these studies adopted a decentralized system architecture.
2. Algorithm-based models were used slightly more than the mathematical models by the reviewed papers.
3. In a related finding, the results showed that to conduct their experiments, about 44% of the studied papers implement their ad hoc simulations. This might be an indication that existing frameworks are not meeting all the needs required for implementing UAV simulations.

In this thesis, we adopt both the concepts of goal-driven cognitive agent architecture, the BDI model and decentralized system architecture when proposing our architecture (*see* Chapter 6). This is formulated in the research question 5.2.3 (page 68). Additionally, the proposed model is algorithm-based (*see* Section 6.4.2). For the ABS implementation, we have two implementations. The first one is in the pilot test (Chapter 8) where we use Repast Simphony [68] based on a comparison we have conducted (*see* Appendix B). In the second implementation in the main test (Chapter 9), we use an ad hoc simulation like most of the studied papers in the SLR.

Chapter 5 provides a detailed discussion of the research methodology conducted in the thesis and it details the research questions of the thesis. Additionally, it explores the role of ABS in the experimental methodology of the proposal of the thesis.

[5] *see* Appendix A for SLRQ4, SLRQ5, SLRQ6, SLRQ7, and SLRQ8.

III

CONTRIBUTION

5

RESEARCH METHODOLOGY

5.1/ INTRODUCTION

The goal of this chapter is threefold: *(i)* to give more details on the general problem of this thesis by summarizing the context, the technological and scientific problems and the related works done in the previous chapters; *(ii)* to set out the features, Research Questions (RQs) and Research Hypotheses (RHs) on which the contribution is based; *(iii)* to outline the experimental methodology conventionally conducted for the validation and evaluation of the contribution in the domain of Explainable Artificial Intelligence (XAI).

With the widespread use of Artificial Intelligence (AI) systems, 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 the agent's state of mind. Recent works in the literature highlighted explainability as one of the cornerstones for building trustworthy and responsible AI systems [77, 176, 233, 244]. Consequently, the sub-domain research of XAI gained momentum both in academia and industry [113, 16, 51].

From Chapter 4, recent studies in goal-driven systems, *e.g.* robots and agents, have confirmed that explaining the system's behavior to the humans fosters the latter's trust in the system and increases its acceptability (refer to Section 2.4 for more details on goal-driven systems). The problem of understanding the behavior of robots is more accentuated in the case of remote robots since —as confirmed by recent studies in the literature [124, 23]— remote robots tend to instill less trust than robots that are co-located. Additionally, providing overwhelming or unnecessary information may also confuse the humans and cause failure.

Despite considerable advances, the domain of XAI is still in its early stages of development, and we can consolidate the general problem of this thesis as **How to build an adaptive context-aware architecture, model, explanation process, and simulation tool to support the human-agent explainability for goal-driven AI systems in the**

context of remote robots (*e.g.* UAVs)?

Remote robots are represented as agents in the thesis (*see* Section 2.3). To achieve smooth human-agent interaction and deliver the best possible explanation to the human, the contribution of the thesis must satisfy the main features of an explanation (refer to Section 3.3.1 for more details):

FEATURE 1: SIMPLICITY

The explanation should be relatively simple to consider the human cognitive load.

FEATURE 2: ADEQUACY

The explanation should include all the pertinent information to help the human understand the situation, even in abnormal situations where the remote robot tends to diverge from the behavior expected by the human.

In Chapter 3, we have outlined how works in the literature have started to respond to these two features. Some works tackled the simplicity of explanation [121, 122, 120, 160], and others investigated the adequacy [145, 60, 202, 241, 302, 214, 56, 268], while few works tried to handle the trade-off between simplicity and adequacy [162]. Additionally, very few works have conducted empirical human studies [185] to properly evaluate the contributions. We have discussed these related works identifying their models and approaches along with analyzing their results and in particular their shortcomings (refer to Section 3.3.1 for more details). Despite considerable advances, the domain of XAI is still in its early stages of development, and there are some open research issues to be tackled. Therefore, we have discussed and summarized the open research issues in the domain of XAI in general and the goal-driven XAI in particular. We have organized these issues into four categories to facilitate identifying the RQs that this thesis handles (refer to Section 3.3.3 for more details about the open research issues). The next section discusses more details about the RQs.

5.2/ RESEARCH QUESTIONS

We have chosen the RQs from each of the first three categories (agent typology, parsimony, and architectures and models) we have identified for the open research issues in the literature (*see* Section 3.3.3) keeping in mind that the RQs should be related and consistent with each other. For the fourth category (evaluation), we opt to consider it in the empirical human studies. Namely the following RQs are investigated.

5.2.1/ EXPLAINABILITY FOR REMOTE AGENTS

The first RQ to tackle concerns the benefits of using explainability in the domain of remote robots. In Chapter 3, explainability has been confirmed to be useful to increase the humans' understandability of robots and agents in various domains, but not trust. This thesis focuses on the domain of goal-driven systems (*see* Section 2.4) and in particular on remote robots, *e.g.* Unmanned Aerial Vehicles (UAVs). Therefore, we need to reproduce and validate the results from the literature in this specific domain. RQ1 can be formulated as follows:

Research Question 1

Does explainability increase the humans' understandability of the remote robots represented as agents?

5.2.2/ PARSIMONIOUS EXPLANATIONS

The second RQ investigates the trade-off between simplicity and adequacy. Even though some few works tried to tackle this question (*e.g.* [162, 214]), the results were not decisive because the model did not consider the human cognitive load. Additionally, the significance of the results obtained in the empirical human study conducted in these works were questionable, *e.g.* one work included only 17 participants [162]. We argue that the parsimony of explanation (*see* Section 2.5) is one of the key characteristics allowing successful human-agent interaction with a parsimonious explanation to be the simplest (*i.e.* least complex) explanation 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 at the conceptual aspect, and without a conjunctive solution that combines the phases of providing an explanation (*see* Section 2.7). RQ2 can be formulated as follows:

Research Question 2

How to strike a balance between simplicity and adequacy?

RQ2 could be divided into two sub-questions RQ2-1 and RQ2-2 defined as follows:

Research Question 2-1

How to provide a simple explanation without the risk of oversimplification?

The goal of this sub is to investigate the level of simplicity beyond which the explanation is oversimplified, and what is the mechanism to guarantee to respect this level.

Research Question 2-2

How to ensure all the necessary information are included in the explanation without overwhelming the human?

The goal of this sub-RQ is to study the process of building the explanation in both normal and abnormal situations. Additionally, it investigates whether this is performed in a single step by the remote agents in the generation phase of an explanation or several steps in the generation and communication phases of the explanation.

5.2.3/ Modeling Explainability for Remote Agents using Cognitive Architectures

Even though we have seen different models and architectures to model explainability, there was no way to compare which one suits best the explainability and XAI. In the literature of goal-driven XAI (Chapter 3), it has been shown that the most used social science theory is folk psychology on which cognitive architectures rely. Moreover, besides ad hoc models, the most used model is the Belief-Desire-Intention (BDI) model to generate explanations for goal-driven agents. Additionally in the literature of Agent-based Simulation (ABS) in UAVs (*see* Chapter 4), most of the works have used cognitive architectures (*see* SLRQ1 on page 43). The question here is: Can we rely on cognitive architectures and models as good candidates for human-agent explainability for remote agents?

In this question, we do not try to compare architectures and models to find out the best candidate to model explainability for remote agents. We simply reproduce the choices of the literature to confirm if a cognitive architecture and a BDI model provide good candidates for human-agent explainability in the application of remote robots represented as agents. RQ3 can be formulated as follows:

Research Question 3

Are the cognitive architecture and the BDI model good candidates for human-agent explainability?

5.2.4/ RESPONDING TO THE RESEARCH QUESTIONS

To well answer the RQs, the thesis proposes a mechanism for parsimonious XAI. In particular, it introduces the process of explanation formulation and proposes a human-agent explainability architecture, named HAExA (*see* Section 2.8), allowing to make it operational for agents. HAExA considers the three phases of providing an explanation from agents to the human: generation, communication, and reception (*see* Section 2.7). We argue that a well-formed adaptive and context-aware combination of these phases leads to formulating a parsimonious explanation. In particular, the proposed architecture relies on a formulation process that includes generating contrastive explanations (*see* Section 2.6) and communicating filtered explanations.

Chapter 6 mainly provides the contributions to answer the RQs. RQ1 is mainly handled by the proposed architecture HAExA namely in Sections 6.2 and 6.3, and the part of the proposal related to RQ1 is evaluated in the pilot test (Chapter 8) to prove that it answers this RQ. We argue that a parsimonious explanation helps to answer RQ2, which is thoroughly discussed in Section 6.5. In particular, Section 6.5.1 investigates the building of explanations and Section 6.5.2 investigates the communication and filtering of explanations. Finally, Sections 6.2 and 6.4 investigate RQ3. The parts of the proposal related to RQ2 and RQ3 are evaluated in the main test (Chapter 9) to prove it answers these RQs.

5.3/ RESEARCH HYPOTHESES

In the process of evaluating the proposal, the RQs should be consolidated into RHs that could be statistically analyzed for significance. Therefore, we pose the following RHs based on the RQs defined before. Before stating the RHs, we need to define some concepts that will be used in the RHs.

In human-agent interactions, there are two types of roles for the human. The first role is “in-the-loop” where the human is involved in the processes of the environment. The other role is “on-the-loop” where the human is less involved, perhaps best described as a supervisor [92]. The human-on-the-loop role is defined in Definition 15 by Nahavandi [211], while Definition 16 provides our version of this role merging parts and components from the literature. We opt to adopt this role for the human in our proposal because we are interested in the effect of explainability on the understandability and trust of the human, *i.e.* no active involvement of the human is reacquired. Additionally, an based on various sources from the literature, we define the adaptive filtering of explanations in Definition 17.

Definition 15: Human-on-the-loop according to Nahavandi [211]

The role of the human where the machines can execute a task completely and independently but have a human in a monitoring or supervisory role, with the ability to interfere if the machine fails.

Definition 16: Human-on-the-loop in explainability

A human whose 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.

Definition 17: Adaptive filtering of explanations

The filtering of explanations in the run-time based on some aspect, *e.g.* the context or the user model.

Accordingly, the thesis considers the following RHs:

RH1 Explainability increases the understandability of the human-on-the-loop in the context of remote agents[1].

RH2 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.

RH3 This research hypothesis could be divided into three sub-hypotheses:

- **RH3-1** **Adaptive filtering** with only normal explanations increases the understandability of the human-on-the-loop compared to **static filtering** with only normal explanations.
- **RH3-2** **Adaptive filtering** with normal and **contrastive** explanations, *i.e. parsimony of explanations*, increases the understandability of the human-on-the-loop compared to **static filtering** with only normal explanations.
- **RH3-3** **Adaptive filtering** with normal and **contrastive** explanations, *i.e. parsimony of explanations*, increases the understandability of the human-on-the-loop compared to **adaptive filtering** with only normal explanations.

RH4 **Adaptive filtering** with normal and **contrastive** explanations, *i.e. parsimony of explanations*, increases the trust of the human-on-the-loop compared to **static filtering** with only normal explanations.

[1] Remote agents represent the remote robots.

According to Mefteh et al. [194], for a MAS with an adaptive behavior [103], the empirical approach is probably the only possible approach in the validation activities. Therefore, to accept or reject these RHs, a specific experimental methodology should be conducted that includes empirical human studies. This methodology considers the recommendations and requirements needed to test any contribution in the XAI domain. The following section discusses thoroughly this experimental methodology.

5.4/ Experimental Methodology

5.4.1/ Agent-based Simulation for Realizing Empirical Human Studies

One important objective of research on explainable agents is the evaluation of explanation approaches in Human-Computer Interaction (HCI) studies[2]. Considering that the growth of research on explainable agents is accelerating, contributions that empirically evaluate the proposed explainability approaches are still scarce [15, 198]. In this regard, ABS (*see* Section 2.9 for more details) fits the requirements to implement such empirical evaluations (refer to Section 2.9 for more details). ABS can be considered as a natural step forward towards better managing and evaluating the proposed explainability approaches in empirical human studies. To facilitate more research and bridge the gap between the theoretically proposed explainability approaches on the one hand, and the practical evaluation of such approaches on the other hand, this thesis presents an ABS approach to engineer explainable agents and Multi-agent System (MAS) prototypes for the specific purpose of empirical evaluation in human studies.

We employ ABS to provide a proof of concept of the contributions as an application of UAVs that is designed to assess the effect of different explainability approaches on the human intelligibility of explanations. Chapter 7 provides the experimental case study, where we conduct an empirical human study to evaluate HAExA with its models and processes by investigating several RHs related to the human-agent explainability. The study is about the delivery of packages using civilian UAVs as remote robots represented by agents. The choice of ABS is based on these reasons:

(i) It allows reproducing the complexity of the behavior of remote agents and the collective behavior.

(ii) Unlike tests with real robots, simulated robots do not need access to expensive hardware and field tests that are costly, time-consuming, and require trained and

[2]named “human studies” in the rest of this thesis

skilled people. Moreover, in the field, it may also be difficult to reproduce the same scenario several times [179].

The human study is performed in two tests. In the pilot test (Chapter 8), the implementation includes only reactive agents. We rely on the agent framework Repast Simphony [68] that controls and manages the environment and the scheduler of the agents. However, when performing the main test (Chapter 9), we opt to develop an ad hoc solution based on the JS-son agent-oriented programming library [139, 206]. In the main test, we make use of light-weight web technologies that facilitate rapid prototyping and allow for the deployment of agents and MAS as static web pages. We present a way to facilitate human studies by implementing explainable agents and MAS that *(i)* can be deployed as static files, not requiring the execution of server-side code, which minimizes administration and operation overhead, and *(ii)* can be embedded into web front ends and other JavaScript-enabled user interfaces, hence increasing the ability to reach a broad range of human users.

The participants in the study watch the simulation execution and then fill out a questionnaire built according to the XAI metrics in the literature [132]. The results of the questionnaire are used to investigate the human understandability and trust of the explanations provided by the UAVs. The next section discusses in detail the proper way to statistically analyze the participants' responses.

5.4.2/ Statistical Testing

To perform the tests, the thesis focuses on qualitative data. The two most common types of qualitative data (nominal and ordinal) are used in the test. The nominal data refer 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 [171]. The writing of the choices of responses may differ in some questions but the scale is the same. However, although the Likert scale is widely used in scientific research, there has been a long-standing controversy regarding the analysis of ordinal data [269]. Analyzing the outcomes of the Likert scale, and the use of parametric tests to analyze ordinal data in general, has been subject to an active and ongoing debate. There are two main opposite points of view of researchers:

- **Likert scale is compatible with parametric testing:** According to these researchers [218, 35, 269], the analysis of the Likert scale can be performed with parametric tests such as ANOVA, t-test, *etc*. As asserted by Norman [218]: "parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of coming to the wrong conclusion". The main arguments of these researchers are:

  - The analysis of the Likert scale by parametric tests such as ANOVA is common in literature. Therefore, if parametric tests cannot be used for the analysis of the Likert scale data, then almost 75% of the research on various domains should be discarded [218].
  - Likert scale data is ordered and can be converted to interval data, *i.e.* to quantitative data [35], which means that parametric tests can be used.
  - Parametric tests can be used with samples of small sizes Norman [218], and this claim is supported by the literature of statistics.
  - ANOVA can be used on non-normally distributed data [218] since the condition of its applicability is not the normal distribution of data but the normal distribution of means.

- **Likert scales are not compatible with parametric testing:** According to these researchers [163, 136, 67], the analysis of Likert scales must be done with non-parametric tests such as Kruskal-Wallis or Mann-Whitney [163]. This is justified as follows:

  - The assumption that Likert scales constitute interval level measurement, which is qualified as ”confusing” by this group of researchers who consider that converting Likert scale to interval data is incorrect [163, 136]. Notably because the response categories in Likert scales have a rank order, but the intervals between values cannot be presumed equal. For example, Cohen et al. [67] contend that it is illegitimate to infer that the intensity of feelings between strongly disagree and disagree is equivalent to the intensity of feelings between other consecutive categories on the Likert scale.
  - The mean, standard deviation, variances are inappropriate for ordinal data [136]. To this end and according to these researchers, median or mode suit more ordinal data to measure the central tendency of the data.

For the pilot test (Chapter 8), a non-parametric test which is Mann-Whitney U is used considering that the pilot test is just a preparation for the main test. However for the main test (Chapter 9) and to avoid biases in the data analysis and due to the dispute between researchers and statisticians, the methodology adopted is to conduct both the parametric test that is ANOVA and the non-parametric test that is Kruskal-Wallis for the data analysis.

5.5/ CONCLUSION

The research methodology conducted in the thesis is five-fold:

(i) Identify the general problem and the main features that should be respected when providing an explanation to the human. Then analyze the related work accordingly to identify the open research issues.

(ii) Define the RQs based on the open research issues.

(iii) Structure the RQs in RHs that can be statistically analyzed.

(iv) Propose the architecture, the model, and the process to answer the RQs.

(v) Conduct a specific experimental methodology to evaluate the proposals by statistically investigating the RHs according to the recommendations in the XAI domain.

Figure 5.1 depicts this research methodology. The boxes of contributions, human study using ABS, and statistical testing are going to be investigated in the next chapters (Chapters 6, 7, 8, and 9). Accordingly, the final version of the research methodology is provided in Chapter 10 (Figure 10.1 on page 144).

To well answer the RQs, the thesis proposes a mechanism for parsimonious XAI. In particular, it introduces the process of explanation formulation and proposes HAExA a human-agent architecture allowing to make this process operational for remote robots which are

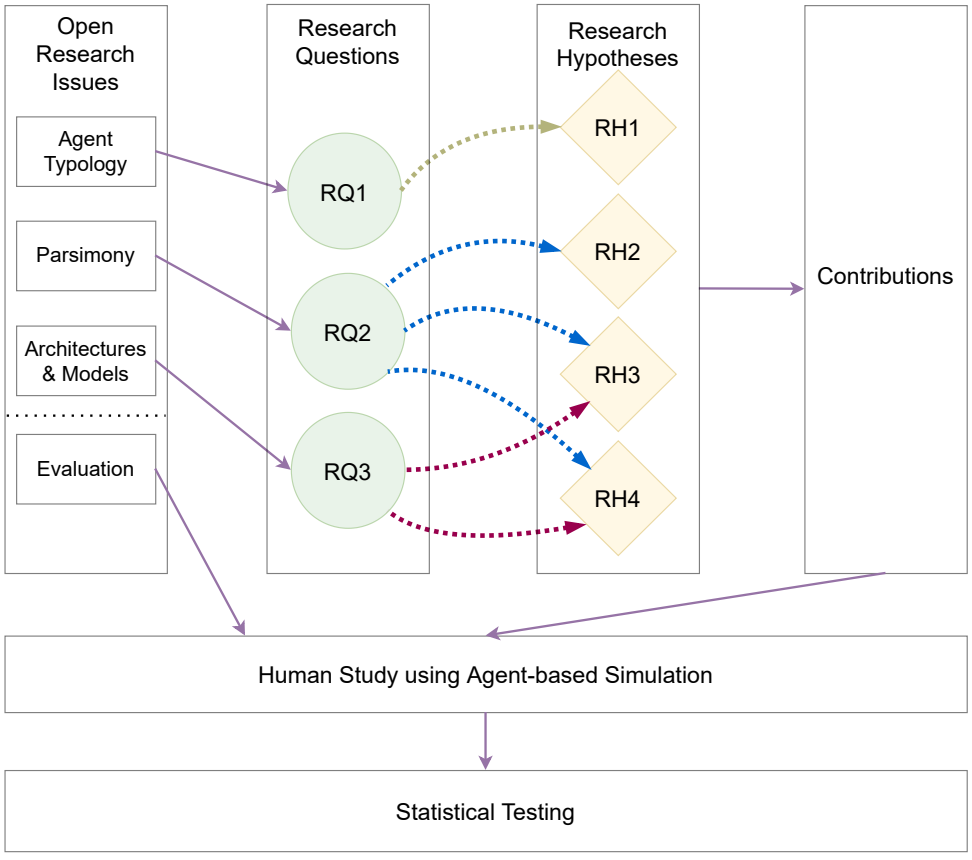


Figure 5.1: Research methodology of the thesis (version-1)

represented as BDI agents (*see* Section 2.8).

The proposed architecture investigates the three phases of providing an explanation from agents to the human: generation, communication, and reception (*see* Section 2.7). We argue that a well-formed adaptive and context-aware combination of these phases leads to formulating a parsimonious explanation. To achieve this, the proposed architecture relies on a formulation process that includes generating contrastive explanations and communicating filtered explanations. Chapter 6 mainly provides the contributions to answer the RQs as follows:

- **RQ1:** It is mainly handled by the proposed architecture HAExA namely in Sections 6.2 (page 78) and 6.3 (page 79).
- **RQ2:** We argue that a parsimonious explanation help answering RQ2 which is thoroughly discussed in Section 6.5 (page 90). In particular, Section 6.5.1 investigates the generation of explanations and Section 6.5.2 investigates the communication and filtering of explanations.
- **RQ3:** It is investigated in Sections 6.2 (page 78) and 6.4 (page 82).

The part of the proposal related to RQ1 is evaluated in the pilot test (Chapter 8), while the parts related to RQ2 and RQ3 are evaluated in the main test (Chapter 9).

To evaluate the proposal, the RHs are investigated in empirical human studies. To conduct these studies, ABS tools (*see* Section 2.9) are developed to implement a proof of concept of the proposed architecture, model, and process. These tools should facilitate the subjective evaluation of the explanation approaches in the proposal by humans participating in the evaluation process. Additionally, the studies rely on well-established XAI metrics and questionnaires (*see* Section 2.10) in the literature to estimate how trustworthy and satisfactory the explanations provided by HAExA are for humans. Like for any empirical human study, the significance of the results should be statistically analyzed using parametric and non-parametric statistical testing.

The next chapter proposes the theoretical contribution of the thesis including the human-agent explainability architecture HAExA, the BDI-based model of the agents, and the explanation formulation process.

6

HUMAN-AGENT EXPLAINABILITY ARCHITECTURE (HAEXA)

6.1/ INTRODUCTION

This chapter describes the contributions of this thesis. These contributions tackle the Research Questions (RQs) and the Research Hypotheses (RHs) defined in Chapter 5. They are threefold:

i) Propose HAExA, an agent-based architecture that facilitates the human-agent explainability, where remote robots are represented as agents. HAExA helps in facilitating the formulation of the necessary explanations communicated from the remote agents to humans, while at the same time considering the human cognitive load to avoid overwhelming him/her with too many details in the explanation.

ii) Define a Belief-Desire-Intention (BDI) based model of the remote agents that generate the explanations and the assistant agent that communicate these explanations. Both the generation and communication of explanations are based on the beliefs and intentions of the agents.

iii) Propose an adaptive context-aware process of explanation formulation based on the parsimony of explanations (*see* Section 2.5). The latter uses various combinations of generating and communicating the explanations.

This chapter is organized as follows. First, Section 6.2 outlines the definitions and general principles of HAExA. Second, Section 6.3 offers a detailed overview of the agents and their roles in HAExA. Third, Section 6.4 extends the BDI model for the agents in HAExA. In particular, the agent practical reasoning cycle is discussed and analyzed in detail. Fourth, Section 6.5 proposes the explanation formulation process. The latter includes the generation and the communication of explanations. Finally, Section 6.6 concludes this chapter.

6.2/ Definitions and General Principles of HAExA

As discussed in the previous sections, explanations are formulated to take into account the properties of the underlying Artificial Intelligence (AI) systems, the context, the features of the explanation (simplicity and adequacy), and the cognitive load of the human who receives the explanations. Therefore, we define the explanation formulation process in Definition 18.

Definition 18: Explanation Formulation Process

A process that seeks to maximize the explanation's adequacy concerning an AI system while minimizing its impact on the human's cognitive load, *i.e.* maximizing its simplicity.

To operationalize the explanation formulation process to a wide range of human-agent interactions, we introduce the Human-Agent Explainability Architecture (HAExA). This architecture allows remote robots, represented as agents and organized in a Multi-agent System (MAS), to expressively explain their behaviors in various situations to humans. The human in HAExA is considered as a human-on-the-loop[1] (*see* Definition 15 on page 70). This is justified by the fact that the human has a passive role in the environment, *i.e.* the human receives explanations for after-action decisions without altering the processes in the environment, *i.e.* no active involvement in the environment is required from the human.

From Definition 18, the explanation formulation process aims to strike a balance between the adequacy, *i.e.* the informativeness of the explanations, and simplicity, *i.e.* to consider the limited human cognitive load (refer to Section 3.3 for more details). To implement and operationalize this process, HAExA proposes a dynamic approach to integrate the three phases of an explanation, *i.e.* generation, communication, and reception [214] (refer to Section 2.7 for more details). 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 with respect to the various situations they face. The way these raw explanations are generated in HAExA varies according to the explained behavior or the situation, either normal or abnormal. Normal explanations are generated in normal situations and contrastive explanations (*see* Section 2.6) in abnormal situations. 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

[1] For the sake of conciseness, the term 'human' is henceforth used to refer to the human-on-the-loop.

humans. Regardless of the approach, the main important goal is to provide explanations that include all the useful information and that are intelligible to humans.

2 - Explanation Communication in HAExA : This step is handled by an assistant agent positioned in between the remote agents on the one hand and the human on the other hand. It is responsible for assuring two tasks: *(i) Update* the raw 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* them; This will facilitate a better understanding by the human, notably because the communicating agent receives the raw explanations from all the remote agents in the MAS. Therefore, it holds 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 agent 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. Better reception of explanations could be also achieved by building a user model to understand the preferences of the human. However, the latter is out of the context of this thesis even though HAExA permits the modeling of the explanation reception phase. Instead, we focus, in HAExA, on how the internal states of remote agents are aggregated and processed to finally be presented as explanations. Therefore, this phase is considered by the empirical human studies conducted later in the thesis, where Agent-based Simulation (ABS) is used to facilitate the reception of explanations by humans based on the recommendations and metrics in the XAI domain.

Explanation formulation involves the three mentioned phases in a way that allows for conducting them separately or in conjunction to provide the human with parsimonious explanations that attain the two features of simplicity and adequacy. Considering the mentioned phases and principals of providing explanations, the following section offers a detailed overview of the agents and their roles in HAExA.

6.3/ AGENTS IN HAEXA

Figure 6.1 visualizes HAExA that is composed of three different entities:

- 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

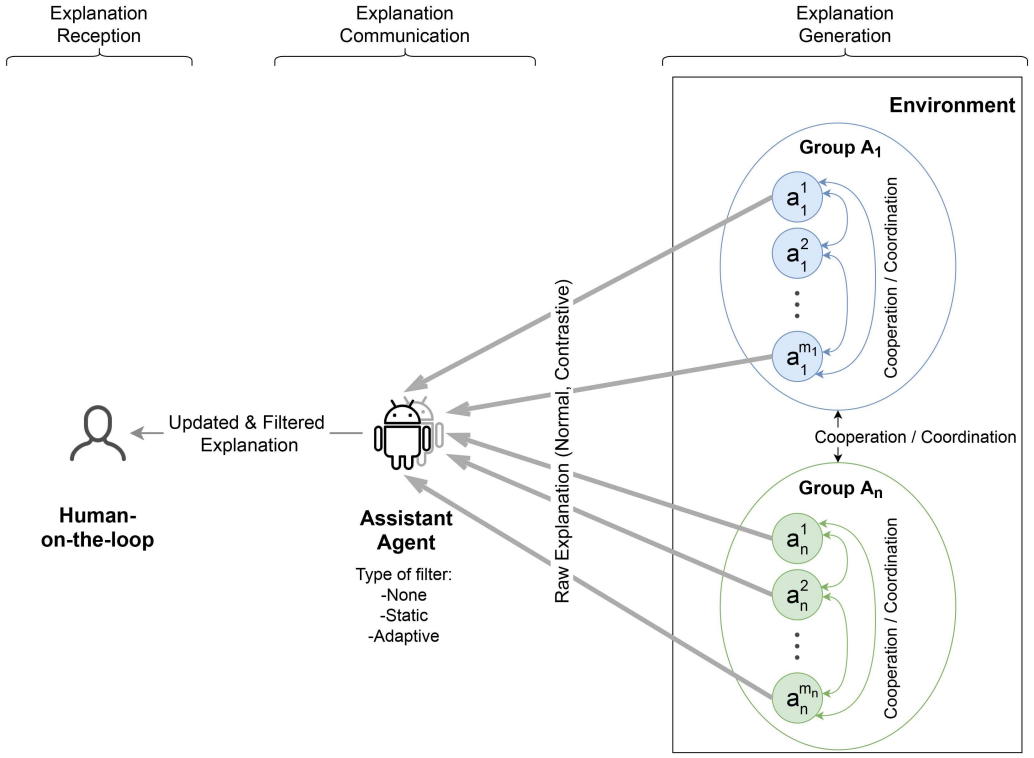


Figure 6.1: Human-Agent Explainability Architecture (HAExA)

the scalability of the architecture. Both in and across groups, collaboration and coordination among agents may occur, while competition and malicious behavior of agents are out of the scope of the thesis. 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.

- An **assistant agent** (illustrated in the center of figure) that collects the remote agents' raw explanations. Then, and considering that the assistant agent has a global overview of the environment, it may update the raw explanations received from the remote agents to guarantee their adequacy. Additionally, it filters the raw explanations before communicating them to the human, as humans could easily get overwhelmed by the information the remote agents provide [273]; subsequently, it communicates the 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).

The reason for choosing an architecture of 3 layers is to well integrate it within the 3 phases of an explanation, namely generation, communication, and reception (*see* Section 2.7). If the number of remote agents scales largely, the layer where the explanations are generated, *i.e.* the remote agents can be sub-divided into sub-layers to achieve scalability in large-scale situations. Some very recent works have started to tackle the issue

of scalability of BDI agents [277, 276]. The use of Holonic MAS [99] could be employed to extend HAExA in this direction by representing every group, *e.g.* a swarm of UAVs, as a holon that has one representative (the head). Koestler [154] coined the term holon as an attempt to conciliate holistic and reductionist visions of the world. A holon represents a part-whole construct that can be seen as a component of a higher-level system or as a whole composed of other self-similar holons as substructures depending on the situation or the perspective. This also applies to the assistant agents that could be formed in a group for fault tolerance and backup reasons and to avoid having a bottleneck in the architecture. Additionally, holons inside an assistant agent holon may execute different filtering behaviors and interact/cooperate to provide the explanation to the human. The most important point is to have one interface with the human user to avoid overwhelming her/him with many interfaces. We can look at the assistant agent as the personal assistant of the human that could be embedded in its smartphone for example. Therefore, and even with a group of assistant agents, the interface with the human is preferably unified through one agent, and the concept of holons is a good modeling candidate for that thanks to the "Janus effect" [154].

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 remote agents formed as in Equation 6.1:

$$A = \{A_1, A_2, ..., A_n\} \tag{6.1}$$

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 the remote agents. Let's assume a_i^j the j^{th} remote agent of the group i. This group of remote agents can be defined as in Equation 6.2.

$$A_i = \{a_i^1, a_i^2, ..., a_i^{m_i}\} \tag{6.2}$$

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

HAExA is flexible and compatible with different agent architectures used by remote agents and the assistant agent. In this thesis, we implement HAExA twice: First, using reactive agents in the pilot test (Chapter 8); Second, using BDI agents in the main test (Chapter 9). However, we opt here to continue explaining the various aspects and processes of HAExA with the BDI cognitive architecture for the reasons mentioned in Section 2.8. As a quick reminder of these reasons, the BDI cognitive architecture resembles the manner humans think because it is based on the concepts of folk-psychology [63, 37]. Therefore,

it has been outlined as a good candidate to represent everyday explanations [217] since it is viewed as the attribution of human behavior using *'everyday'* terms such as beliefs, desires, intentions, emotions, and personality traits [63, 186, 45]. Moreover, the BDI cognitive architecture allows agents to exhibit more complex behavior than purely reactive architectures but without the computational overhead of other cognitive architectures [5]. Furthermore, some evidence exists that BDI agent architectures facilitate knowledge elicitation from domain experts [85]. Finally, BDI has been identified as the most used architecture to generate explanations for goal-driven agents (*e.g.* [45, 214]) [16].

6.4/ The Proposed Belief-Desire-Intention Model

6.4.1/ General Principles

The BDI model is a model of human behavior that was developed by philosophers. The BDI model 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 [41]. The conceptual framework of the BDI model is described in [42].

For pedagogical reasons, we shortly describe the different concepts of beliefs, desires, and intentions of the BDI model as follows [37]:

- **Beliefs:** Information that the agent has about the environment and may be out of date or inaccurate, *e.g.* the locations of a package to be delivered.
- **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.
- **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 select 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.

Detailed implementations of the coordination and cooperation among agents are out of the scope of the thesis, as these aspects are already covered in-depth by a range of research works. The reader is referred to [309] for more details about the definitions and main characteristics of cooperation and coordination.

6.4.2/ AGENT PRACTICAL REASONING CYCLE

The particular model of decision-making in the BDI model is known as practical reasoning. The latter is a process directed towards actions from the notations of beliefs, desires, and intentions [316]. Considering that HAExA provides explanations to humans, there is a need to understand how humans do the reasoning process. Human practical reasoning consists of two distinct activities:

- Deliberation which is fixating upon options that the human wants to achieve, *i.e.* going from desires to intentions;
- Means-ends reasoning which is deciding how to act to achieve these intentions using the available means or actions [37]. Means-ends reasoning is generally known in the AI community as *planning* [104].

The practical reasoning cycle of a remote agent includes the internal processing performed within the agent to act in the environment and coordinate with the other agents based on three aspects: *(i)* the perceptions from the environment; *(ii)* the messages received from the other agents; *(iii)* its internal beliefs, desires, and intentions. Additionally, and apart from the actions of the agent in the environment, the explanations will be another output of the reasoning cycle.

In HAExA, Algorithm 1 outlines, in pseudo-code, how these concepts are outlined theoretically (*i.e.* the agent control loop) for a remote agent. This algorithm is mainly based on the work of Wooldridge [316, 317] and Bordini et al. [37]. It is adapting the BDI algorithm proposed by these works by adding modules to facilitate the human-agent explainability. For pedagogical reasons, we restate the main algorithm and summarize its components after adding our contributions related to explainability. These contributions are minimized to simple functions in the algorithm for clarity reasons. However, they are thoroughly discussed in Section 6.5.

In Algorithm 1 of a remote agent, the variables are defined as follows:

- B_0 and I_0: The initial beliefs and intentions, respectively, of the agent at the beginning of the execution;
- π: The current plan adopted by the agent;
- B_{Old} and I_{Old}: The previous beliefs and intentions, respectively, of the agent from the previous control cycle. These variables store the current beliefs and intentions before updating them into new ones;
- π_{Old}: The previous old plan adopted by the agent;

Algorithm 1: The control loop of a remote BDI agent (adapted from [316, 317, 37])

Input: B_0: Initial beliefs
Input: I_0: Initial intentions

$B \leftarrow B_0$ $I \leftarrow I_0$;
$\pi \leftarrow$ *null*;
while true **do**
  $B_{Old} \leftarrow B$;
  $I_{Old} \leftarrow I$;
  $\pi_{Old} \leftarrow \pi$;
  $P \leftarrow$ *getPerceptions*();
  $M \leftarrow$ *receiveMessages*();
  $B \leftarrow$ *updateBeliefs*(B, P, M);
  $D \leftarrow$ *updateDesires*(B, I);
  $I \leftarrow$ *updateIntentions*(B, D, I);
  $\pi \leftarrow$ *plan*(B, I, Ac);
  while not (*empty*(π) or *succeeded*(I, B) or *impossible*(I, B)) **do**
    $a \leftarrow$ *head*(π);
    executeAction(a);
    sendMessages(B, I, π);
    if *abnormalSituation*(B, B_{Old}, I, I_{Old}) **then** *contrastiveExp*(B, I, π, π_{Old}) ;
    else *normalExp*(B, I, a) ;
    $\pi \leftarrow$ *tail*(π);
  $P \leftarrow$ *getPerceptions*();
  $M \leftarrow$ *receiveMessages*();
  $B \leftarrow$ *updateBeliefs*(B, P, M);
  if *reconsider*(I, B) **then**
    $D \leftarrow$ *updateDesires*(B, I);
    $I \leftarrow$ *updateIntentions*(B, D, I);
  if not *sound*(π, I, B) **then**
    $\pi \leftarrow$ *plan*(B, I, Ac);

- P: The current perceptions of the agent about the environment received via its sensors;
- M: The messages received from other agents in the MAS;
- B: The current beliefs of the agent that represent information it has about its environment;
- D: The current desires of the agent that are the options it is considering, *i.e.* are candidates to be intentions;
- I: The current intentions of the agent that contain the states of affairs it has chosen and committed to achieving;
- Ac: The full list of all possible actions the agent can perform based on its actuators;

- a: One action of the adopted plan;

The functions in Algorithm 1 are defined as follows:

- $getPerceptions()$: The remote agent observes its environment to get the next perception. One of the things that can be done using a simulated environment like the one used in the tests in this thesis is to determine the properties of the environment which are only perceivable by one particular agent or more, *i.e.* an individualized perception can be defined. This is useful because all the remote agents have individualized perceptions while the assistant agent has a global view thanks to the explanations and messages received from all remote agents.

- $receiveMessages()$: The agent receives messages from other agents in the MAS. Mainly, these messages are for cooperation to perform missions and assign tasks.

- $updateBeliefs(B, p)$: The agent updates its beliefs about the environment that reflects the changes in the environment. Updating the beliefs is achieved in the following way. Considering P is the set of current perceptions and B is the set of literals in the beliefs that were obtained from sensing the environment:

  1. each literal l in P not currently in B is added to B;

  2. each literal l in B no longer in P is deleted from B.

- $updateDesires(B, I)$: The agent determines its desires, or options, based on its current beliefs and intentions.

- $updateIntentions(B, D, I)$: The agent chooses between its desires and selects some to become intentions. The main discarded desires are the unrealistic ones or the ones that are impossible to fulfill based on the current beliefs.

- $plan(B, I, Ac)$: The agent generates a plan to achieve its current intentions based on the actions it can perform. The inner loop from line 13 to line 19 shows the execution of this plan. The agent considers one action in turn from the plan π and executes it until the plan is empty, *i.e.* all the actions in the plan have been executed. After executing an action, the agent explains it based on its beliefs and desires.

- $empty(\pi)$: This function simply checks if a list or set (here, the plan) is empty.

- $succeeded(I, B)$: This function checks if the current intentions are achieved according to the current beliefs.

- $impossible(I, B)$: This function checks if the current intentions are impossible to achieve based on the current beliefs.

- $head(\pi)$: This function simply takes the first element of a list or set (here, the plan).
- $executeAction(a)$: The agent executes an atomic action directly possible via its actuators.
- $sendMessages(B, I, \pi)$: The agent sends messages to other agents in the MAS. Mainly, these messages are based on the beliefs, the intentions, and the plan of the agent.
- $abnormalSituation(B, B_{Old}, I, I_{Old})$: This function verifies if the situation the agent is facing is either normal or abnormal. For that, it compares the current new beliefs and intentions with the old beliefs and intentions, respectively, from the previous cycle. then it checks if the difference is higher than a specific threshold for beliefs and a specific threshold for intentions. If both differences are above the corresponding thresholds, the situation is abnormal from the perspective of the agent (refer to Section 6.5.1 for more details).
- $contrastiveExp(B, I, \pi, \pi_{Old})$: This function generates a contrastive explanation based on the current beliefs, current intentions, the current plan to be performed, and the old plan from the previous cycle (refer to Section 6.5.1.2 for more details).
- $normalExp(B, I, a)$: This function generates a normal explanation based on the current beliefs, current intentions, and the atomic action to be executed (refer to Section 6.5.1.1 for more details).
- $tail(\pi)$: This function simply returns all the elements of a list or a set (here, the plan) without the first one.
- $reconsider(I, B)$: After executing an action from the plan (line 15), the agent pauses to perceive the environment again to update its beliefs. Then, it checks, using this function, if it is worth reconsidering its intentions (*i.e.* spending time deliberating over them again). This decision is based on the idea that this reconsideration may lead to changing the intentions. Otherwise, it is better not to waste time and computational efforts in deliberation, and it is better to continue trying to achieve the intentions [148, 251, 252].
- $sound(\pi, I, B)$: The agent verifies whether or not the plan it currently has is sound with respect to its intentions and beliefs. If it believes the plan is no longer a sound one, then it chooses a new plan.

Figure 6.2 depicts the practical reasoning cycle of a remote agent based on the generic BDI architecture [238]. The figure outlines the internal processing performed within the agent to act in the environment and coordinate with the other agents. In this figure,

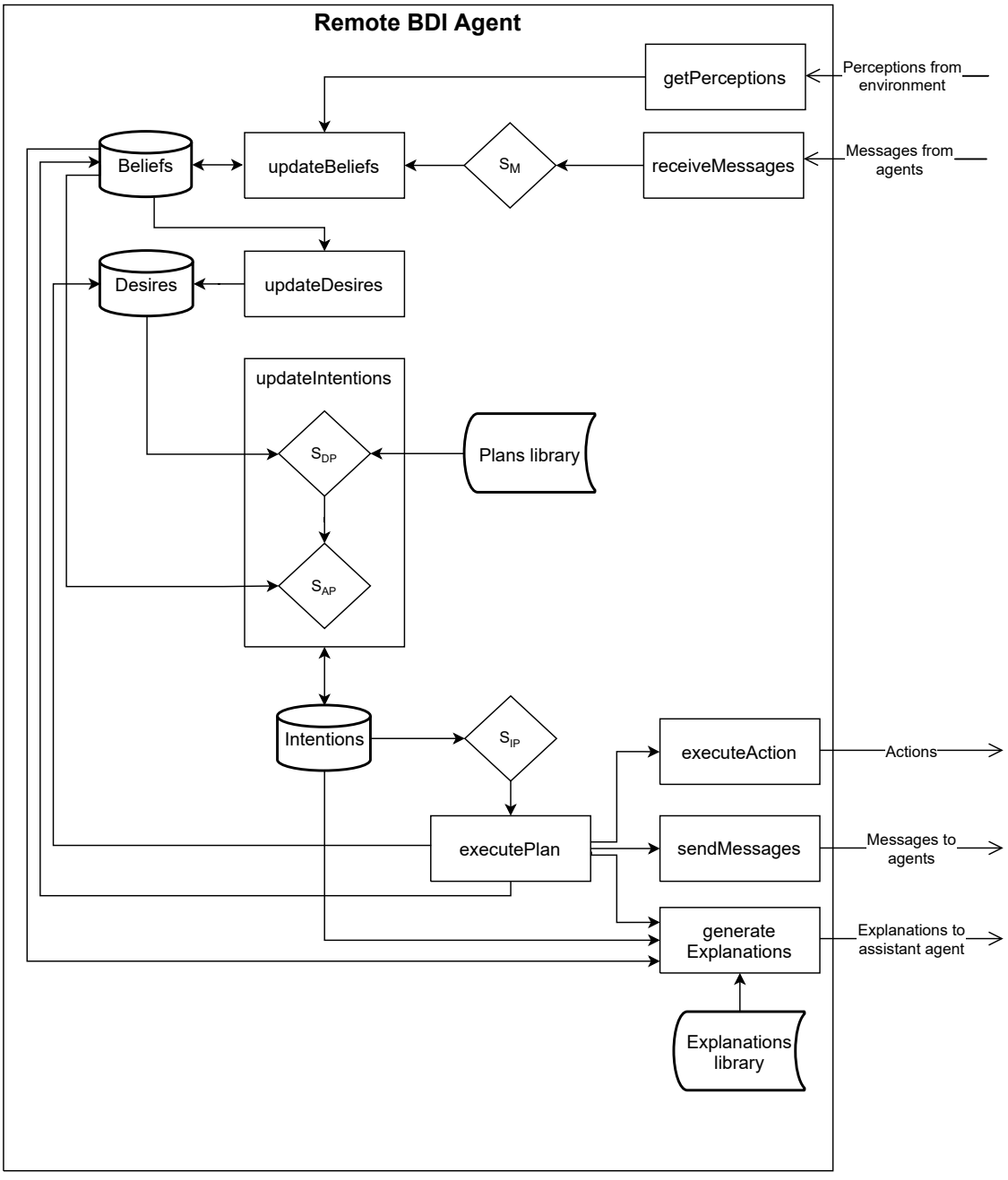


Figure 6.2: The practical reasoning cycle of a remote BDI agent

cylindrical shapes represent the main architectural components that determine the state of the agent, *i.e. Beliefs*, *Desires*, *Intentions*. The trapezoid shapes (or data storage shapes) store useful data in the design-time like the *Plans library* and *Explanations library*. The former includes all the actions (Ac in Algorithm 1 on page 84) that can be used in a plan. The latter includes components of explanations as phrases to be used when building the explanations. Rectangles (*getPerceptions, receiveMessages, updateBeliefs, updateDesires, updateIntentions, executePlan, executeAction, sendMessages, generateExplanations*) represent the main functions executed in the reasoning cycle (Also can

be found in Algorithm 1). Diamonds represent the selection functions used to select an item from different choices.

The *MAS communication* module, consisting of the $receiveMessages$ function, the $sendMessages$ function and the S_M (selectMessages) selection function in Figure 6.2 is responsible for organizing the communication with other agents by managing the messages received from the other agents in the MAS and the messages to send. It also allows for the exchange of information and know-how, and the delegation of intentions. In some situations, the agent would like to filter messages coming from untrusted agents. Additionally, in many situations, agents would like to give priorities to certain messages and that is why we use the S_M selection function. Several mechanisms could be used to facilitate the communication between the agents. HAExA is open to any possibility but we could recommend some particular mechanisms, *e.g.* Human-Agent-Robot-Machine-Sensor (HARMS) [189, 300], Knowledge Query and Manipulation Language (KQML) [91], Web Ontology Language (OWL) [192], or Semantics of Business Vocabulary and Business Rules (SBVR) [107, 108] to cite a few. In particular, HARMS connects actors over a network by a peer-to-peer manner and uses particular message types such that all actors are indistinguishable in terms of which type of actor (*e.g.* robot, software agent, or even human) sends a message [190]. However, as all agents in HAExA only cooperate to serve the human there is no mistrust between the agents, and hence only the priority of messaging is considered. Therefore, the way how this module is implemented will not be further discussed.

Apart from the S_M selection function, all the other selection functions used in Figure 6.2. These selection functions have the following functionalities:

- S_{DP} (selectDesirablePlans): This selection function is part of the *updateIntentions* function (Figure 6.2). It is responsible for retrieving all the *desirable plans* that allow the agent to act to fulfill its desires. This selection is governed by the desires of the agent and the components (actions or sub-plans) available in the *Plans library*, *i.e.* this function selects the intentions that can be represented as plans and can fulfill the desires at the same time (refer to [316, 317] for more details about the formality of plans).

- S_{AP} (selectApplicablePlans): This selection function is part of the *updateIntentions* function (Figure 6.2). It is used because not all the desirable plans can be excused, as some of them may be not realizable considering the current beliefs. Therefore, this selection function is used to select the applicable plans that can be executed. For that, we need to verify whether the context of each of the desirable plans is believed to be true, *i.e.* whether the context is a logical consequence of the beliefs of the agent.

- S_{IP} (selectIntentionalPlans): At this step of the reasoning cycle, there are several applicable plans that are candidates for execution to achieve an intention and hence fulfill a desire. However, the agent needs to choose only one to execute. The first idea to come to mind is to go with a random choice. The second is to use a scheduling mechanism, *e.g. Round-Robin* [168, 226]. Other more sophisticated run-time mechanisms could be defined. HAExA is context-aware to the abnormality of the situation which is used to generate context-aware explanations. This mechanism could also be used to prioritize the intentional plans to be executed, *e.g.* giving high priority to those who are applicable in abnormal situations where the abnormality is determined based on the difference between the current beliefs and/or intentions on the one hand and the previous ones on the other hand.

The *generateExplanations* module is responsible for either generating raw normal explanations in normal situations or raw contrastive explanations in abnormal situations. The generation is based on the beliefs and intentions of the agent on the one hand and explanation components from the *Explanations library* on the other hand. Full details about this module are offered in Section 6.5.1.

To reduce redundancy, the practical reasoning cycle of the assistant agent is not provided. The two main differences with the practical reasoning cycle of the remote agent are:

i) The assistant agent is not interacting with the environment. Therefore, it does not get perceptions nor execute actions. Instead, it aggregates the beliefs received by all remote agents to form his own beliefs. Then, it follows the same practical reasoning like the remote agent (Figure 6.2) to update its own desires and intentions. In the case of conflicting beliefs received from different remote agents, ad hoc solutions are employed to resolve the conflicts. However, HAExA is open for advanced mechanisms and algorithms of aggregation.

ii) The explainability module in the assistant agent is different as instead of generating explanations, the assistant agent receives raw explanations generated by the remote agents (normal and contrastive), and post-processes these explanations before communicating them to the human. The post-processing of the raw explanations includes: First, updating the type of the explanation (normal or contrastive) as the assistant agent has a global view of the system, and hence its beliefs are the complete ones about the situation. Second, filtering the updated explanations. Full details about the explainability module of the assistant agent are provided in Section 6.5.2.

In the next section, we investigate with full details the generation and updating of the two types of explanations (normal and contrastive) and the different mechanisms of filtering.

6.5/ 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 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. In particular, we focus in this section, as an instance of the process of providing explanations using HAExA, on generating normal and contrastive explanations as well as using filtering of explanations for communicating the explanations. Figure 6.3 shows the process of the explanation formulation pipelines. For the generation, two distinct methods are considered: normal explanations ($normalExp(B, I)$ in Algorithm 1 on page 84) in normal situations (Section 6.5.1.1), and contrastive explanations ($contrastiveExp(B, I)$ in Algorithm 1) in abnormal situations (Section 6.5.1.2). For the communication, three means of filtering are considered: static filter, adaptive filter, and no filter (investigated in Section 6.5.2.2). Additionally, and before filtering the explanations, there is a sub-step of updating the raw explanations received from the remote agents (investigated in Section 6.5.2.1). This instance of the explanation formulation process is used in the implementation of the main test (Chapter 9) later in the thesis.

Mainly, the rest of this section is organized as follows. Section 6.5.1 discusses the generation of explanations by the remote agents. Section 6.5.2 tackles the communication of explanations by the assistant agent.

6.5.1/ Explanation Generation

All remote agents provide to the assistant agent the set of all raw explanations $RExp$ that can be, based on Equations 6.1 and 6.2 (page 81), represented in Equation 6.3.

$$RExp = \bigcup_{i=1}^{n} \bigcup_{j=1}^{m_i} rExp_i^j \tag{6.3}$$

where $n \in \mathbb{N}^*$ is the number of the groups of the 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 is based on 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

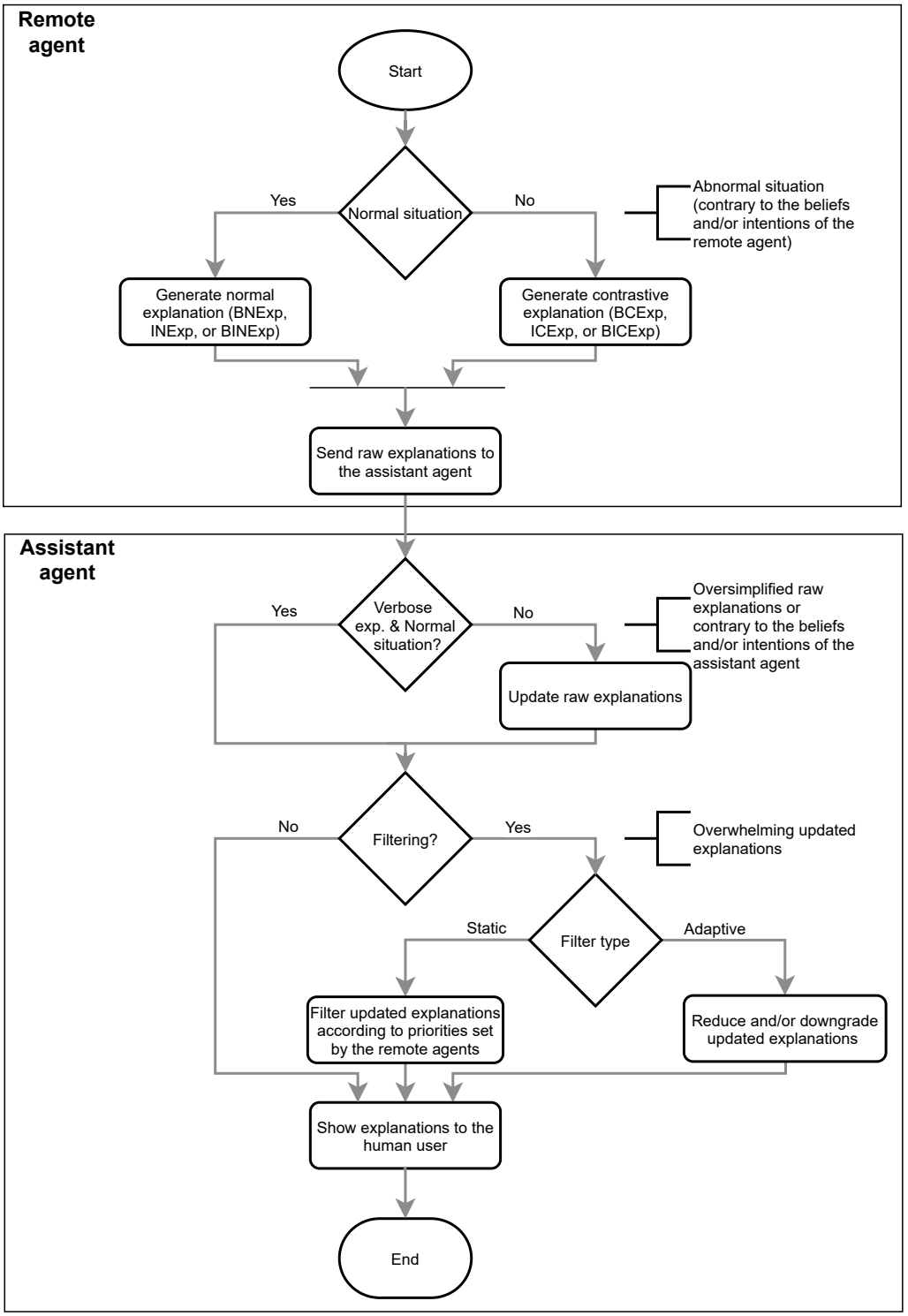


Figure 6.3: HAExA explanation formulation process

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.* $B - B_{Old}$, and if there were beliefs previously present but disappeared in the updated current beliefs, *i.e.* $B_{Old} - B$. Δ_B is defined

in Equation 6.4. Accordingly, the condition to generate $RExp$ based on the beliefs is defined in Equation 6.5.

$$\Delta_B = reduceRedundancy(B \cup B_{Old} - B \cap B_{Old}) = reduceRedundancy((B - B_{Old}) + (B_{Old} - B)) \tag{6.4}$$

where the function $reduceRedundancy$ eliminates all redundant attributes.

$$\Delta_B > \theta_{Belief} \tag{6.5}$$

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 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 6.6). Accordingly, the condition to generate $RExp$ based on the intentions is defined in Equation 6.7.

$$\Delta_I = reduceRedundancy(I \cup I_{Old} - I \cap I_{Old}) = reduceRedundancy((I - I_{Old}) + (I_{Old} - I)) \tag{6.6}$$

$$\Delta_I > \theta_{Intention} \tag{6.7}$$

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 explanations $RExp$ (Normal or Contrastive) by the remote agents, and then maybe updated by the assistant agent. These remote agents are BDI agents, whose beliefs and intentions are used to generate $RExp$. As stated before, a raw explanation $rExp_i^{m_i}$ could be of two types: normal explanation in normal situations and contrastive explanation in abnormal situations. Both types are further divided into sub-types as described in the next sections.

6.5.1.1/ NORMAL EXPLANATIONS

Generally, in normal situations, the remote agent generates the explanation 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: *normal* explanation. It is stating the next step in the 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 delivering 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 law battery", "UAV 1 is charging battery", *etc*.

Figure 6.4 shows the structure of an Intention Hierarchy Tree (IHT) which is based on Goal Hierarchy Trees (*see* Section 3.3.1.1). A remote BDI agent acts and explains its actions based on the IHT as follows. Based on the current intentions and beliefs of the agent, it chooses an action to perform. If multiple actions are applicable, then it randomly chooses one unless an explicit priority between actions is defined. When no actions are applicable then it remains on standby until its beliefs change according to the changes in the environment, which can cause it to commit to new intentions, and then new actions could become applicable. The IHT (Figure 6.4) can be used to determine Δ_B and Δ_I, *e.g.* we consider that there is a significant difference between I and I_{Old} if there are n levels of differences in the IHT where $n \in [1..r]$ and r is the number of levels of the IHT.

Accordingly, normal explanations could be divided into three sub-types:

- *Belief-based Normal Explanation (BNExp):* In this type, the explanation is based on the current beliefs of the agent. This type has been used before in the literature (*cf.* Section 3.3.1.1). It mainly states the beliefs above (one level or more) the action to perform in the IHT (Figure 6.4), *e.g.* Action 1 is explained by Belief B1. *BNExp* is generated by a function in Equation 6.8.

$$BeliefRawNormalExp : B \times Ac \rightarrow RExp \tag{6.8}$$

  where B is the set of current beliefs, Ac is the set of all available actions, and $RExp$ is the set of raw explanations.

- *Intention-based Normal Explanation (INExp):* In this type, the explanations are based on the current indentations (or goals) of the agent. This type has been used before in the literature (*cf.* Section 3.3.1.1). It mainly states the intentions above (one level or more) the action to perform in the IHT (Figure 6.4), *e.g.* Action 1 is explained by Intention B. *INExp* is generated by a function in Equation 6.9.

$$IntentionRawNormalExp : I \times Ac \rightarrow RExp \tag{6.9}$$

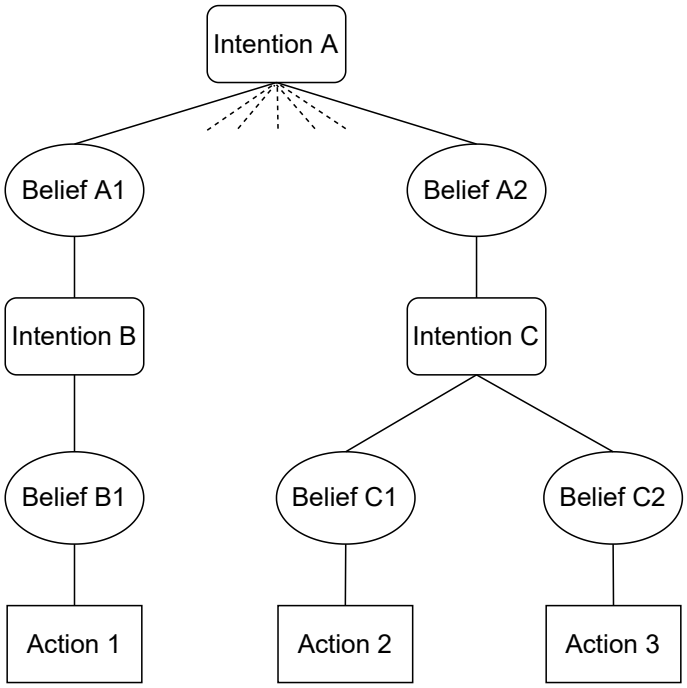


Figure 6.4: Intention Hierarchy Tree (adopted from [250])

where I is the set of current intentions, Ac is the set of all available actions, and $RExp$ is the set of raw explanations.

- *Belief & Intention based Normal Explanation (BINExp):* In this type, the explanations are based on both the current beliefs and the current indentations (or goals) of the agent. It mainly states the beliefs and intentions above (one level or more) the action to perform in the IHT (Figure 6.4), *e.g.* Action 1 is explained by Belief B1 and Intention B. *BINExp* is generated by a function in Equation 6.10.

$$BeliefIntentionRawNormalExp : B \times I \times Ac \rightarrow RExp \tag{6.10}$$

where B is the set of current beliefs, I is the set of current intentions, Ac is the set of all available actions, and $RExp$ is the set of raw explanations.

The function $normalExp(B, I, a)$ defined in Algorithm 1 (page 84) could execute one of the three functions defined in Equations 6.8, 6.9, and 6.10. This depends mainly on the values of the change in beliefs Δ_B (Equation 6.4) and the change in intentions Δ_I (Equation 6.6). Accordingly, the choice of which function to execute depends on threshold values as follows:

1. $\epsilon_{BeliefRawNormal}$ as an instance of θ_{Belief} defined in Equation 6.5 for choosing the function $BeliefRawNormalExp$;

2. $\epsilon_{IntentionRawNormal}$ as an instance of $\theta_{Intention}$ defined in Equation 6.7 for choosing the

function $IntentionRawNormalExp$;

3. $\epsilon_{BeliefCombinedRawNormal}$ and $\epsilon_{IntentionCombinedRawNormal}$ as instances of, respectively, θ_{Belief} defined in Equation 6.5 and $\theta_{Intention}$ defined in Equation 6.7 for choosing the function $BeliefIntentionRawNormalExp$.

It is important to note here the relation between these defined thresholds, as triggering one will prevent triggering the others. These relations are defined in Equations 6.11 and 6.12.

$$\epsilon_{BeliefCombinedRawNormal} > \epsilon_{BeliefRawNormal} \tag{6.11}$$

$$\epsilon_{IntentionCombinedRawNormal} > \epsilon_{IntentionRawNormal} \tag{6.12}$$

The point is that we want to assure that the function $BeliefIntentionRawNormalExp$ is executed if there are changes both in beliefs and intentions because this happens when the situation starts to become abnormal and hence more information is needed to be provided to the human.

If both the thresholds $\epsilon_{BeliefRawNormal}$ and $\epsilon_{IntentionRawNormal}$ are attained, the agent could randomly choose what function to execute (either $BeliefRawNormalExp$ or $IntentionRawNormalExp$). Alternatively, we could rely on the differences between the changes in levels in the IHT.

6.5.1.2/ CONTRASTIVE EXPLANATIONS

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 accordingly, the situation is considered abnormal. In such situations, the contrastive explanations are preferable [198]. Let the following example that is leading to an abnormal situation when a UAV is moving to a package with the intention to deliver it. However the package is delivered by another UAV, the following two explanations of the atomic actions $move$ could be provided: *(i)* "UAV 1 is moving to Package 1"; *(ii)* "UAV 1 is moving to Package 2". In this situation, the human may ask "why did this happen?" and "why UAV 1 did not carry out the delivery of Package 1?" The human will be curious about knowing why the UAV 1 did not do what it was supposed to do, and instead, it is moving to another package. To solve this, an alternative contrastive explanation is provided as follows: "UAV 1 is moving to Package 2 instead of Package 1 because Package 1 is delivered by another UAV".

This contrastive explanation is explaining the action $a1$: "UAV 1 is moving to Package 2", which is part of the new plan π instead of the action $a2$: "UAV 1 is moving to Package 1", which is part of the old plan π_{Old} from the previous reasoning cycle based on the current belief $b1$: "Package 1 is delivered by another UAV".

For the human, and when receiving explanations about the behavior of the remote agents, it is generally not the normal behavior that may be appealing to receive an explanation for, but rather the abnormal behavior. Therefore, we adopt in HAExA the contrastive explanations to represent the abnormal situations. For the human, normal behaviors can be explained with the help of his/her own experiences and expectations. However, the abnormal behavior of the agent challenges these experiences and expectations, and therefore, an explanation is deemed necessary in this case. In abnormal situations, the deviation from the chosen plan will lead to a significant update of the beliefs of the remote agent and accordingly updating the intentions significantly. In such situations, a contrastive explanation is generated.

There could be several options for generating contrastive explanations where $a1$ is an action from the new updated plan π and $a2$ is an action from the old plan 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 it did not (line 6 in Algorithm 1 on page 84). 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 transforming a contrastive explanation into a normal one by dropping the part "not $A2$". This transformation is appealing to reduce the length of the explanation when it can be implicitly inferred by the human. This is another aspect of the adaptivity of the model to achieve parsimonious explanations where the trade-off between simplicity in normal explanations and adequacy in contrastive explanations is achieved. This is all based on the context and the human cognitive load. To do that, the assistant agent has a role in the updating of the generated raw explanations, as it holds a general comprehensive overview of the context situation; hence it may aggregate more useful and consistent explanations for the human by updating $RExp$ generated by the remote agents. Accordingly, normal explanations by some remote agents could be filtered and contrastive explanations by others could be kept if there is a need to update the quality of the explanations for adequacy reasons or reduce their quantity for simplicity reasons. The updating and filtering of explanations are discussed in detail in Section 6.5.2.

As C could be either an intention or a belief or both, there are three sub-types of contrastive explanations: Belief-based Contrastive Explanation (BCExp), Intention-based

Contrastive Explanation (ICExp) and Belief & Intention based Contrastive Explanation (BICExp) that are generated as functions in Equations 6.13, 6.14, and 6.15 respectively.

$$BeliefRawContrastiveExp : B \times Ac \times Ac \rightarrow RExp \tag{6.13}$$

$$IntentionRawContrastiveExp : I \times Ac \times Ac \rightarrow RExp \tag{6.14}$$

$$BeliefIntentionRawContrastiveExp : B \times I \times Ac \times Ac \rightarrow RExp \tag{6.15}$$

where B is the set of current beliefs, I is the set of current intentions, Ac is the set of all available actions, and $RExp$ is the set of raw explanations. Ac is used twice because the input includes two elements from Ac: the new action to perform and the old action that was not performed.

It is intuitive to notice that the BICExp is more adequate and less simple than BCExp and ICExp. This will be useful in the adaptive communication of the filtered explanations as it will help in balancing the trade-off between simplicity and adequacy. Further details are provided in the next section.

The function $contrastiveExp(B, I, \pi, \pi_{Old})$ defined in Algorithm 1 (page 84) could execute one of the three functions defined in Equations 6.13, 6.14, and 6.15. This depends mainly on the values of the change in beliefs Δ_B (Equation 6.4) and the change in intentions Δ_I (Equation 6.6). Accordingly, the choice of which function to execute depends on threshold values as follows:

1. $\epsilon_{BeliefRawContrastive}$ as an instance of θ_{Belief} defined in Equation 6.5 for choosing the function $BeliefRawContrastiveExp$;
2. $\epsilon_{IntentionRawContrastive}$ as an instance of $\theta_{Intention}$ defined in Equation 6.7 for choosing the function $IntentionRawContrastiveExp$;
3. $\epsilon_{BeliefCombinedRawContrastive}$ and $\epsilon_{IntentionCombinedRawContrastive}$ as instances of, respectively, θ_{Belief} defined in Equation 6.5 and $\theta_{Intention}$ defined in Equation 6.7 for choosing the function $BeliefIntentionRawContrastiveExp$.

It is important to note here the relation between these defined thresholds for contrastive explanations, and the threshold of generating normal explanations $\epsilon_{BeliefRawNormal}$, $\epsilon_{IntentionRawNormal}$, $\epsilon_{BeliefCombinedRawNormal}$, and $\epsilon_{IntentionCombinedRawNormal}$ as triggering one will prevent triggering the others. These relations are defined in Equations 6.16 and 6.17, which are updated versions of Equations 6.11 and 6.12 on page 95.

$$\epsilon_{BeliefCombinedRawContrastive} > \epsilon_{BeliefRawContrastive} > \epsilon_{BeliefCombinedRawNormal} > \epsilon_{BeliefRawNormal} \tag{6.16}$$

$$\epsilon_{IntentionCombinedRawContrastive} > \epsilon_{IntentionRawContrastive} > \epsilon_{IntentionCombinedRawNormal} > \epsilon_{IntentionRawNormal} \tag{6.17}$$

The point is that we want to assure that the functions to generate contrastive explanations are checked for execution before the functions of the normal explanations when there are significant changes both in beliefs and intentions. This happens when the situation is abnormal, and hence more information is needed to be provided to the human. If both the thresholds $\epsilon_{BeliefRawContrastive}$ and $\epsilon_{IntentionRawContrastive}$ are attained, the agent could randomly choose what function to execute (either $BeliefRawContrastiveExp$ or $IntentionRawContrastiveExp$). Alternatively, we could rely on the differences between the changes in levels in the IHT.

6.5.2/ Communicating Updated and Filtered Explanations

This section discusses how the assistant agent communicates the explanations to the human. This step is divided into two sub-steps:

1. Updating $RExp$ (Section 6.5.2.1) that takes $RExp$ generated by the remote agents as input and output *updated explanations* as defined by a function in Equation 6.18.

$$updateExp : RExp \rightarrow UExp \tag{6.18}$$

   where $UExp$ is the set of updated explanations.

2. Adaptive filtering of $UExp$ (Section 6.5.2.2) that takes $UExp$ as input and outputs the *filtered explanations* to the human as defined in Equation 6.19.

$$filterExp : UExp \rightarrow FExp \tag{6.19}$$

   where $FExp$ is the set of filtered explanations to be communicated to the human.

These two sub-steps will be discussed in detail in the next two sections.

6.5.2.1/ Updating the Raw Explanations

The raw explanations $RExp$, generated by the remote agents, are updated by the assistant agent to assure they are adequate, *i.e.* they hold all the necessary information. This sub-step of the explanation formulation process is important to assure that the generation of explanations did not result in the oversimplification of the explanations. The results

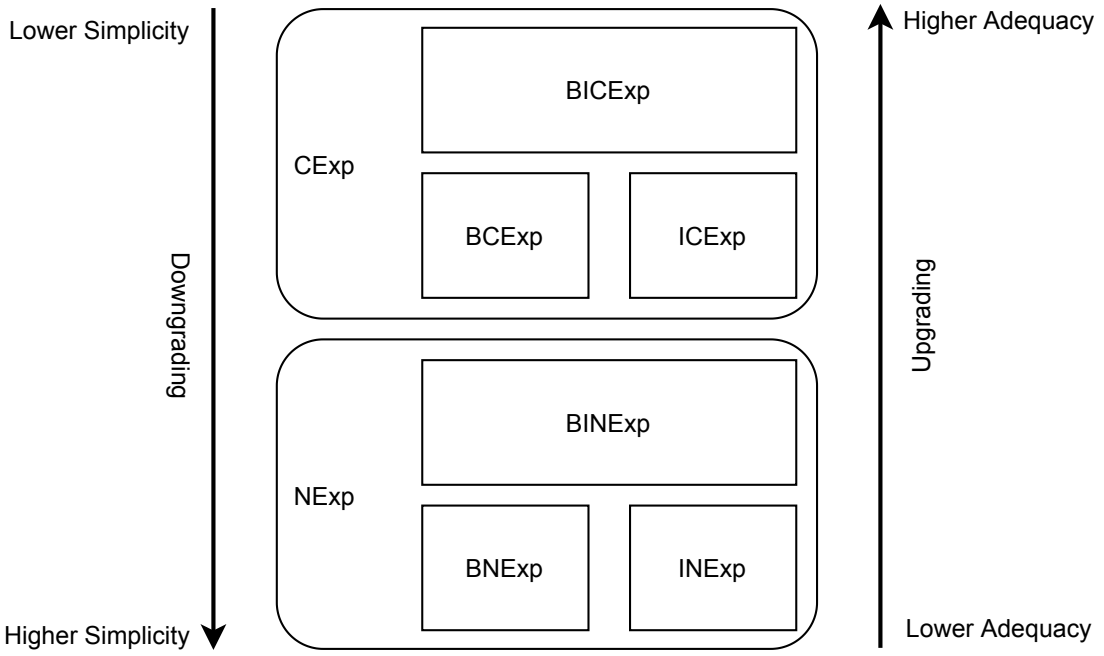


Figure 6.5: The hierarchy of the explanations levels

of this step are the *updated explanations* $UExp$. This sub-step is context-aware to the situation, *i.e.* it adaptively updates $RExp$ based on the context of the situation. Additionally, 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 are 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, and after receiving all $RExp$, could discover some abnormality hidden to the remote agents. Therefore, if the situations are considered abnormal according to the assistant agent, it updates $RExp$ generated by the remote agents.

For updating, we have defined a hierarchy of *levels of explanations* to adaptively handle the trade-off between simplicity and adequacy. Figure 6.5 depicts this hierarchy where the higher levels having higher adequacy and lower simplicity while the lower levels having higher simplicity and lower adequacy. Accordingly, the hierarchy of explanations could be divided into 4 levels:

1. Belief & Intention based Contrastive Explanation (BICExp): This explanation is the longest but includes the most information;

2. Belief-based Contrastive Explanation (BCExp) and Intention-based Contrastive Explanation (ICExp);

3. Belief & Intention based Normal Explanation (BINExp);
4. Belief-based Normal Explanation (BNExp) and Intention-based Normal Explanation (INExp): These explanations are the shortest but may not include all necessary information.

The assistant agent could *upgrade* an explanation to a higher level (Level 1 is the highest) if there is a need for adequacy, *i.e.* there is a risk of oversimplifying the explanation or *downgrade* the explanation into lower levels (Level 4 is the lowest) if there is a need for simplicity, *i.e.* there is a risk of overwhelming the human. In level 4, BNExp and INExp are considered at the same level of simplicity and adequacy. However, studies that are listed by Kaptein et al. [141] show that based on the age of the human (child or adult), these types of explanations could be distinguished. This discussion is also applicable for Level 2 between BCExp and ICExp. Finally, BCExp and ICExp are considered at a higher level than BINExp. This is explained by the fact that contrastive explanations are associated more than normal explanations with abnormal situations that need more adequate information [198], *i.e.* higher adequacy. However, this ordering means also that BINExp is simpler than BCExp and ICExp. This could be explained by the fact that contrastive explanations are generated when facing abnormal situations that are not simple to understand by the human.

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 perspective of the remote agents. Accordingly, the assistant agent upgrades or downgrades $REXp$ based on the abnormality of the situation. This is confirmed according to the *change in beliefs* Δ_B defined in Equation 6.4 (page 92) and the *change in intentions* Δ_I defined in Equation 6.6 (page 92), while considering eight different thresholds. Four of them are instances of θ_{Belief} defined in Equation 6.5 (page 92): $\epsilon_{BeliefUpdateNormal}$, $\epsilon_{BeliefCombinedUpdateNormal}$, $\epsilon_{BeliefUpdateContrastive}$, and $\epsilon_{BeliefCombinedUpdateContrastive}$. The other four are instances of $\theta_{Intention}$ defined in Equation 6.7 (page 92): $\epsilon_{IntentionUpdateNormal}$, $\epsilon_{IntentionCombinedUpdateNormal}$, $\epsilon_{IntentionUpdateContrastive}$, and $\epsilon_{IntentionCombinedUpdateContrastive}$. Like with the generation of $REXp$, the rules defined in Equations 6.20 and 6.21 govern the relation between these thresholds.

$$\epsilon_{BeliefCombinedUpdateContrastive} > \epsilon_{BeliefUpdateContrastive} > \epsilon_{BeliefCombinedUpdateNormal} > \epsilon_{BeliefUpdateNormal} \tag{6.20}$$

$$\epsilon_{IntentionCombinedUpdateContrastive} > \epsilon_{IntentionUpdateContrastive} > \epsilon_{IntentionCombinedUpdateNormal} > \epsilon_{IntentionUpdateNormal} \tag{6.21}$$

An example of upgrading, if the assistant agent receives an explanation of the type ICExp while the thresholds $\epsilon_{BeliefCombinedUpdateContrastive}$, and $\epsilon_{IntentionCombinedUpdateContrastive}$ are

attained, it will upgrade the raw explanation to BICExp.

6.5.2.2/ FILTERING OF THE UPDATED EXPLANATIONS

From Figure 6.5 (page 99), the types of explanations with higher levels (highest is Level 1) provide more information (*i.e.* more adequate) than the other types with lower levels (lowest is Level 4). However, the textual length of the types in the higher levels is longer than those in the lower levels. With longer explanations (*i.e.* less simple), we risk overwhelming the human, and to face this challenge we introduce two solutions:

1. The remote agent is *context-aware* and *adaptive* to the situation in that it chooses which explanation to provide based on the priority of the situation. The latter could be determined based on several ways:

   i) Fixed list of priorities of situations set in the design-time. The problem with this way is the difficulty to anticipate all possible situations;

   ii) Some preset expert reactive rules. The problem with this way is the difficulty to anticipate all possible events for which rules should be defined;

   iii) The *change in beliefs* Δ_B, *i.e.* based on the change in the environment. For that we need a new threshold $\epsilon_{FilterPriority}$ as an instance of θ_{Belief} defined in Equation 6.5 (page 92).

2. The assistant agent is context-aware and adaptive to the situation in that it filters $UExp$ received from the remote agents to insure the human is not overwhelmed. The priority when providing parsimonious explanations is to provide adequate explanations (*i.e.* based on both beliefs and intentions) unless there is a risk to overwhelm the human according to some *overwhelming threshold*, *i.e.* human cognitive load threshold. In the latter case, simple explanations are kept (lower levels) while longer ones are filtered out. It is very important here to state that the overwhelming threshold is not calculated theoretically in HAExA. Even though calculating this value is very useful, it is out of the context of this thesis. Instead, this value is calibrated empirically using several beta tests where human participants are involved.

The filtering of explanations is conducted to assure that $UExp$ are simple and do not overwhelm the human, *i.e.* increase the simplicity. This sub-step is adaptive to the number of explanations provided by the remote agent 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 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 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 set in the design time. This filter is called a *static filter* and is used in the pilot test (Chapter 8) to investigate RH1 (page 70) and RH2 (page 70).

3. Using an adaptive filter based on the current context, where irrelevant explanations are removed; for example, if many remote agents are present in the environment, the assistant agent may decide to aggregate their explanations because it is not possible for a human to 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.

For adaptive filtering, three levels of adaptation are defined:

- $FilterThreshold_H$: If the number of $UExp$ is higher than this threshold, downgrade $UExp$ into lower levels of hierarchy (*see* Figure 6.5 on page 99) using the function $downgrade$ and reduce the number of $UExp$ of the normal types using the function $reduce$. Reducing the number of explanations may lead to fully discarding them. This type is defined in Equation 6.22.

$$|UExp| > FilterThreshold_H \rightarrow reduce(BINExp, BNExp, INExp), downgrade(UExp) \tag{6.22}$$

- $FilterThreshold_M$: If the number of $UExp$ is higher than this threshold, reduce the number of $UExp$ of the normal types using the function $reduce$. This type is defined in Equation 6.23.

$$|UExp| > FilterThreshold_M \rightarrow reduce(BINExp, BNExp, INExp) \tag{6.23}$$

- $FilterThreshold_L$: If the number of $UExp$ is higher than this threshold, downgrade $UExp$ into lower levels of hierarchy (*see* Figure 6.5 on page 99) using the function $downgrade$. This type is defined in Equation 6.24.

$$|UExp| > FilterThreshold_L \rightarrow downgrade(UExp) \tag{6.24}$$

If two explanations have the same sub-type or same hierarchy level and there is a need to discard one of them, the decision is made based on priorities set by the remote agents when generating the explanations. Finally, 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 upgrading $RExp$ into $UExp$ and then later these $UExp$ are going to be downgraded because they are overwhelming. This condition can be found in Figure 6.3 (page 91).

6.6/ CONCLUSION

This chapter proposed the contributions of this thesis. These contributions tackle the RQs and the RHs defined in Chapter 5. The main goal of the contributions is to provide parsimonious explanations to the human that strike a balance between adequacy and simplicity. Achieving adequacy ensures that all the necessary information is included in the explanations while achieving simplicity avoids overwhelming the human with extra cognitive load. The main contributions proposed in this chapter are:

i) A Human-Agent Explainability Architecture (HAExA) representing remote robots as agents that provide explanations to the humans about the environment, their decisions, and their behaviors.

ii) A BDI-based model of the remote agents generating the explanations and the assistant agent communicating them.

iii) An *adaptive* and *context-aware* explanation formulation process based on the parsimony of explanations. This process seeks to maximize the explanation's adequacy while minimizing its impact on the human's cognitive load. To achieve that, it uses various combinations of generating and communicating the explanations.

Section 6.2 highlighted the definitions and general principles of HAExA. Then, it identified how HAExA adopts and adapts the phases of providing an explanation. Section 6.3 proposed HAExA and presented the MAS within it and how the agents are organized with their roles. In principle, HAExA includes two types of agents: *(i)* The remote agents as part of the environment. *(ii)* The assistant agent whose role is to be an interface between the remote agents and the human.

Section 6.4 discussed in detail the BDI-based model used to represent all agents in HAExA. First, Section 6.4.1 indicated the general principles of designing the BDI model and how HAExA approached them. Second, Section 6.4.2 analyzed in detail the practical reasoning cycle of the agents. In particular, an algorithm of the control loop of a

remote BDI agent has been proposed that extends the well-known BDI algorithm with some modules related to explainability.

Section 6.5 proposed and thoroughly explored the *explanation formulation process*. This section was divided into two main subsections. First, Section 6.5.1 covered the generating of *raw explanations* by the remote agents. These raw explanations concern the situations they perceive, the decisions they made, and the actions they perform. They have two main types: *Normal* in relatively normal situations, and *contrastive* in abnormal ones. These remote agents are BDI agents whose beliefs and intentions are used to generate the raw explanations. The types of explanations are further divided into 6 sub-types depending on the source of information in the explanations: *(i)* based on the beliefs of the agent (BNExp, BCExp); *(ii)* based on the intentions of the agent (INExp, ICExp); *(iii)* based on the beliefs and intentions of the agent (BINExp, BICExp). The choice between these sub-types is concluded by the remote agent based on the change in beliefs and intentions, *i.e.* the remote agent is context-aware.

Section 6.5.2 tackled the communication phase of providing an explanation. The assistant agent has a global view of the context thanks to the raw explanations and messages it receives from the remote agents. Accordingly, it adaptively updates the raw explanations based on the changes in its beliefs and intentions. It mainly upgrades or downgrades the raw explanations according to a defined hierarchy of explanations that tackles the trade-off between simplicity and adequacy. Additionally, it filters the updated explanations respecting the thresholds of the human cognitive load. Finally, it communicates the *filtered explanations* to the human. The thresholds mentioned in this chapter are defined empirically, in general, with their values being calibrated with the help of ABS.

To properly evaluate the proposed contributions of human-agent explainability, we conduct two empirical human case studies based on a scenario of package delivery using civilian UAVs. The next chapter presents further details about this scenario.

IV

EVALUATION

7

Application to Civilian Unmanned Aerial Vehicles

7.1/ Introduction

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, Unmanned Aerial Vehicles (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 [208]. 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 exist 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, in the field, it is hard to reproduce the same scenario several times [179]. In this context, the development of computer simulation frameworks that allow transferring real-world scenarios into executable models is highly relevant [20, 88]. However, the simulation frameworks have their drawbacks; in particular, it is impossible to fully reproduce the real environment.

The use of Agent-based Simulation (ABS) frameworks (refer to Section 2.9 for more details about ABS and to Section 5.4.1 for more details on how to employ ABS for human studies) or tools for UAV simulations is gaining more interest in complex civilian applications where coordination and cooperation are necessary [1]. Due to operational costs,

safety concerns, and legal regulations, ABS is commonly used to implement models and conduct tests for UAVs. This has resulted in a range of research and applied works addressing ABS in UAVs [207].

The problem of understanding the robot's state-of-mind is more accentuated in the case of UAVs since —as confirmed by recent studies in the literature [124, 23]— 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, the evaluation of the contributions of the thesis relies on Explainable Artificial Intelligence (XAI), refer to Section 2.5 for more background details, 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, providing explanations about the remote UAV decisions may increase the satisfaction of humans [40] and maintain acceptability of the XAI system [19]. 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 on 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 [138] 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.

Contrastive explanations (refer to Section 2.6 for more background details and Section 6.5.1.2 for more details on the contribution related to contrastive explanations) is vital in ABS, because the human is watching a simulation in which all normal situations are expected by him/her and while watching and having several normal situations, the human will learn more about the normal behavior of the simulation. Consequently, he/she will be interested in any abnormal behavior and will demand an explanation for such behaviors.

This chapter presents the experiment application and scenario used to evaluate the contributions of the thesis. The scenario employs UAVs as an example of remote robots represented as agents. ABS is mainly used to implement the simulation of the scenario. In the simulation, the agents provide explanations of the environment, their decisions, and their behaviors. The explanations are meant for human participants that will watch the simulation execution and assess its functionalities and explainability. Then, they fill out a questionnaire (refer to Section 2.10 for more background details) built according to the XAI metrics in the literature [132]. The participants' responses in the questionnaire should be statistically analyzed as described in Chapter 5 (*see* Section 5.4.2). The experiment scenario is implemented in two human studies: Pilot test (Chapter 8) and Main test (Chapter 9). These tests aim to investigate the research hypotheses (Section 5.3)

and answer the research questions (Section 5.2) posed in Chapter 5.

The chapter is organized as follows. First, Section 7.2 investigates the experiment scenario with the various roles and the interactions among them. Second, Section 7.3 outlines the principles and the categories for building the questionnaire used to collect the responses of the participants in the evaluation. Third, Section 7.4 highlights the process of conducting the experiment. Finally, Section 7.5 concludes this chapter.

7.2/ 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 [209]. In the scenario, one human-on-the-loop operator (*see* Definition 15 on page 70) oversees several UAVs, *i.e.* the remote agents in HAExA, that will provide package delivery services to clients. These UAVs will autonomously conduct tasks and take decisions when needed. Additionally, they need to communicate and discuss with each other and may cooperate to complete a specific task. The UAVs will explain to the Operator Assistant Agent (OAA), *i.e.* the assistant agent in HAExA, the progress of the mission including the abnormal situations along with the decisions made by them. Figure 7.1 shows the interaction between the actors in the proposed scenario. In the following, the steps of the experiment scenario are detailed:

1. When a client puts a request for delivering a package, a notification is sent to the UAVs, so all UAVs are connected with each other and with the OAA using an assumed reliable network.
2. UAVs that are near, with a specific radius, to the package will coordinate to complete the delivery mission. In other words, if a UAV is very far from 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 for several reasons:
    - **Best candidate:** Deciding which UAV will deliver the package according to constraints: actual distance to the package, battery capacity, having other packages in hand, having a mission with a near destination, *etc*.
    - **Long trip:** 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. If the package is picked up by a UAV from a competitor (external events), we have two situations:

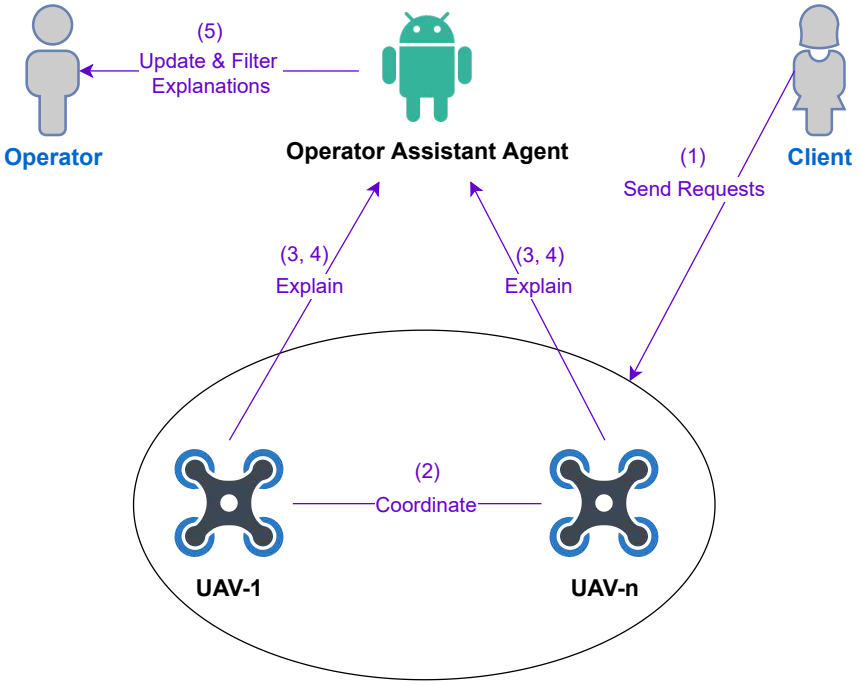


Figure 7.1: The interaction of actors in the experiment scenario

- The client sends a notification that the package/passenger is picked up, and the assigned UAV will stop the mission;
- The client does not send a notification that the package/passenger is picked up (*e.g.* because of selfishness or laziness). In this situation, the UAV will go to the place and observe the absence of the package/passenger, and it needs to explain this to the OAA.

4. The explanation needed from the UAV is generally about the environment, the mission progress, its decisions, its behaviors, its actions, and its status, *e.g.* which UAV is assigned to the mission after the communication between UAVs, or when the UAV picks up the assigned package and is moving to destination. However, other important kinds of explanation are required regarding the abnormal situations, *e.g.* the UAV arrives at the package location and did not find it, or see that it is damaged, or not according to the description (maybe heavier). Another example is when a UAV needs charging and that is why it ignores a nearby package.

5. Every UAV will generate raw explanations for the OAA that will communicate them to the operator. The OAA updates and filters these raw explanations received from the UAVs to give an adequate summary of the most important explanations without overwhelming the operator with a lot of details. There are two main types of explanations: Normal and Contrastive (refer to Section 6.5 for more details on the subtypes of explanations and how to generate and communicate them). There are two types of filtering of explanation by the assistant agent (discussed in detail in Section

6.5.2.2): *(i)* Static filtering based on a filtering threshold set by the human that will filter the explanations based on their priorities. These priorities are set by the UAVs when generating each raw explanation. *(ii)* Adaptive context-aware filtering where the OAA adapts the intensity and levels of filtering based on the complexity of the situation.

The experiment requires that the participants (the operator in the experiment), and after watching the simulation of the experiment, should fill out a questionnaire to collect their opinions on the explanations provided by the agents in the experiment. The next section discusses in detail the parts of the mentioned questionnaire.

7.3/ BUILDING THE QUESTIONNAIRE

We opt to use the Explanation Satisfaction and Trust Scale in building our questionnaire (*see* Section 2.10). The answers are distributed to a 5-points Likert scale [132]:

1	2	3	4	5
I disagree strongly	I disagree somewhat	I'm neutral about it	I agree somewhat	I agree strongly

Our choice is based on two reasons:

1. Unlike the Explanation Quality Checklist [132], it uses a wider $5-7$ scale answers;
2. The Explanation Quality Checklist is intended to be used by researchers who built the XAI system, or as an independent experiment of explanations by other researchers, while the Explanation Satisfaction and Trust scale is mainly suitable for experiments of explanations by humans.

The next section outlines the structure of the built questionnaire with the various categories of the questions.

7.3.1/ CATEGORIES OF THE QUESTIONS

The built questionnaire has 21 questions for the main test and 12 questions for the pilot test 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 that the participant understood the simulation using some objective questions. Additionally, we confirm if the functionalities of the simulation are acceptable by the participant, and their suggestions in this context.

3. *Statistical Analysis (3 questions for the pilot test, see Section 8.4; and 12 questions for the main test, see Table 9.1)* These questions are mainly about the understandability and trust. We investigate the following aspects: satisfaction, confidence, predictability, reliability, efficiency, trustworthiness, and importance of explanation. Additionally, this category specifies questions about explainability to confirm the usefulness of the explanation, satisfaction of the explanation, either or not the details of the explanation have sufficient details, *etc*.

Finally, the questionnaire includes a question (numbered Q18 in the main test) about Curiosity considering that situations like the one understudy lead people to engage in effortful processing and motivates them to seek out additional knowledge to gain insight and fulfill their curiosity [178]. 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 the tests in this thesis do not investigate curiosity, this question is not analyzed in the results of this thesis.

The trust in the automation can rapidly break down under conditions of time pressure, or when there are conspicuous system faults or errors, or when there is a high false alarm rate [82, 184]. The trust of a system can be hard to reestablish once lost. The trust in the automation scale developed by Adams et al. [6] refers specifically to the experiment of simulations [6]. It asks only two questions, one about trust (Do you trust it?) and one about reliance (Are you prepared to rely on it?). The Statistical Analysis category includes the two mentioned questions about Trust along with others in the main test. The next section outlines the steps for conducting the tests that consider the experiment scenario mentioned in this chapter.

7.4/ CONDUCTING THE EXPERIMENT

We have conducted two tests based on the experiment scenario: Pilot test (Chapter 8) and Main test (Chapter 9). In the pilot test, we investigate the research hypotheses RH1 and RH2, while in the main test we investigate the research hypotheses RH3-1, RH3-2, RH3-3, and RH4 (page 70).

The pilot test was conducted in the university with the help of 27 students as participants that were physically present at the university during the test, while the main test was conducted online with 90 participants. The simulation used, for each test, was different as

for the pilot test, we relied on reactive agents, while for the main test the implementation included BDI agents as explained in Chapter 6.

It is important here to mention that before conducting the two tests, all participants have been informed that the gathered data is anonymous, secured, and will be used solely for research purposes. Moreover, they have the right to know how the data is used according to the General Data Protection Regulation[1].

Before conducting the test, we give some information about the simulation: *(i)* Explain the main goal of the simulation, which is the delivery of packages using UAVs. The delivery of a package is from any point of the map, where a package could appear to some warehouse determined on the map. *(ii)* Icons of the elements (UAVs, charging stations, packages, destinations, *etc*.). *(iii)* While the UAVs are delivering the packages, some abnormal situations may happen. All the situations, either normal or abnormal will be explained by the UAVs.

As the coordination and cooperation between the groups of remote agents in the multi-agent system is out of the scope of this thesis, we opt to simplify the implementation of HAExA (*see* Figure 6.1) by choosing only one group.

The simulation is divided into 4 sequences. 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 it is like a happy path situation, which helps 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 will handle an abnormal situation or more, *e.g.* a low battery, a damaged package, an already delivered package, *etc*. The number of abnormal situations increases further with the sequences from the second sequence till the fourth (last) sequence that is an overwhelming sequence with several UAVs (here 10).

7.5/ CONCLUSION

This chapter presented the experiment application and scenario used to evaluate the contributions of the thesis. The scenario considers UAVs as an example of remote robots represented as agents. Different roles for different actors and the interactions among them are identified. ABS will be mainly employed to implement the simulation of the scenario. In the simulation, the agents provide explanations of the environment, their decisions, and their behaviors. The explanations are meant for human participants that will watch the simulation execution and assess its functionalities and explainability. Then, they will fill out an XAI questionnaire based on the Explanation Satisfaction and Trust Scale.

[1] https://gdpr-info.eu/

The questions in the questionnaire are organized into 4 categories, and the responses of the participants are to be statistically analyzed. This section highlighted also the process of experimenting.

We have conducted two tests based on the experiment scenario presented in this chapter: Pilot test (Chapter 8) and Main test (Chapter 9). In the pilot test, we investigate the research hypotheses RH1 and RH2, while in the main test we investigate the research hypotheses RH3-1, RH3-2, RH3-3, and RH4 (page 70). In the following sections, we cover in detail each of these two tests.

8

PILOT TEST

8.1/ INTRODUCTION

In this chapter, we conduct a pilot empirical human study (or pilot test) based on the experiment application (*see* Chapter 7) of package delivery using civilian Unmanned Aerial Vehicles (UAVs) to answer one of the Research Questions (RQs) and evaluate two of the Research Hypotheses (RHs). In particular, the pilot test evaluates the part of the contribution related to RQ1: *Explainability for remote agents* (*see* Section 5.2.1). It tackles and evaluates RH1: *Explainability increases the understandability of the human-on-the-loop in the context of remote agents* and RH2: *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*. (*see* Section 5.3).

In this test, the implementation includes only reactive agents in the Agent-based Simulation (ABS) for realizing the human study (*see* Section 5.4.1). After watching the simulation and filling out the associated questionnaire, the responses undergo statistical testing to interpret, evaluate, and analyze the results (*see* Section 5.4.2).

The chapter is organized as follows. First, Section 8.2 discusses the methodology of conducting the pilot test. Second, Section 8.4 statistically interprets and analyzes the responses of the participants. Third, Section 8.5 outlines the limitations of the test. Finally, Section 8.6 concludes this chapter.

8.2/ PILOT TEST METHODOLOGY

This section conveys the methodology applied to conduct the pilot test as follows. Section 8.3 clarifies some specific details of the implementation and realization of the simulation. Section 8.3.1 details the process of organizing the participants in groups and how to aggregate their responses.

8.3/ EXPERIMENTAL DETAILS

The pilot test discusses the role of filtering of explanations in three cases: No explanation, Detailed explanation, and Filtered explanation. Only normal explanations for normal and abnormal situations are used, *i.e.* no contrastive explanations. Moreover, the normal explanations have only one type based on some preset reactive rules in the design-time, *i.e.* the generation is not based on beliefs nor intentions. Furthermore, only static filtering (*see* Section 6.5.2.2), *i.e.* no adaptive filtering. These simplifications are made because the main goal of this pilot test is to reproduce the results of the literature in the domain of remote robots represented as agents.

The experiment scenario (*see* Section 7.2) in the pilot test is implemented using Repast Simphony [68] as an ABS framework. We rely on this framework to control and manage the environment and the scheduler of the agents. The choice of this framework is based on a comparison of ABS frameworks for unmanned aerial transportation applications showing that Repast Simphony has significant operational and execution features[1]. The agents are all reactive agents and not Belief–Desire–Intention (BDI) agents. This choice was made because, in the pilot test, there is no implementation of adaptive filtering nor contrastive explanations; hence reactive agents are enough to verify the goal of the pilot test which is mainly to reproduce the results of the literature in the domain of UAVs as an example of remote robots, *i.e.* to investigate the impact of filtering the explanations on the understandability of the human in this domain.

The simulation is run 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. Figure 8.1 depicts a snapshot of the simulation presented to the participants for the pilot test [210]. The last sequence of the simulation (overwhelming sequence) lasts for 1:42 minutes and includes: 10 UAVs, 4 warehouses, 5 charging stations, 16 packages to be delivered, 4 abnormal situations. The explanations are textual, and they have a natural language appearance, with the dynamic numbering of the simulation elements (UAVs, packages, charging stations, *etc.*). Some examples of the explanations generated by a UAV agent are: “UAV 1 should carry Package 3” or “Package 4 is damaged. I cannot deliver it”. The UAVs assign priorities to their explanations and the Operator Assistant Agent (OAA) filters the explanations allowing to pass only those with a priority higher than the filtering threshold set by the human in the initial parameters of the simulation.

[1] *see* Appendix B for the latest version of the ABS frameworks comparison.

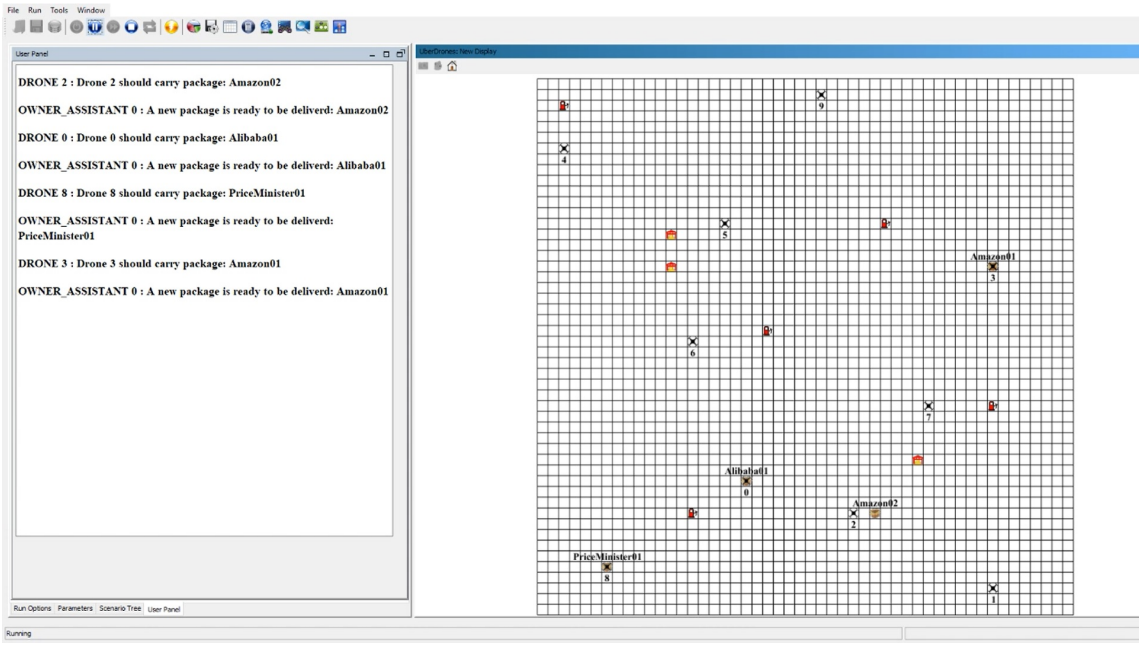


Figure 8.1: Pilot test simulation snapshot

8.3.1/ PARTICIPANTS AND GROUPS

The experiment requires the help of human participants who watch the simulation and then fill out a questionnaire built to aggregate their responses. These responses are distributed to a 5-points Likert scale [132] (refer to Section 7.3 for more details on building the questionnaire). The questionnaire of the pilot test is minimized to 12 questions. Apart from the first 8 questions of Participant Details, Functionalities categories (refer to Section 7.3.1 for more details on the categories of the questions) and the question about curiosity, there are 3 questions left for the statistical analysis on the importance of explanations and the filtering of them (refer to Appendix C for all the questions of the questionnaire in the pilot test). For this test, all participants have the same conditions when watching the execution of the simulation (quality of the video, same place and time, same instructions given, *etc.*). The organizing steps of this test are described as follows:

1. 27 students of the university in the technology domain but in different specialties and different years (Bachelor, Master, and Ph.D.) have participated in this test. They were randomly divided into three groups. Additionally, only the participants with B1 English level have participated. Among the 27 participants, 7 of them were females, 19 were males, and 1 preferred not to disclose this information. They were aged between 21 and 34 (mean of age of the participants is 23.3) with a mean equal to 2.9 (5-points Likert) for self-rated experience with UAVs.

2. All the three groups watch exactly the same simulation sequences but with different explanation capabilities:

1. Group N: 11 participants watch the simulation with no explanation;
2. Group D: 8 participants watch the simulation with detailed textual explanations;
3. Group F: 8 participants watch the simulation with filtered textual explanations.

3. Only one group is allowed to stay in the room at a specific time. The test took 60 minutes: the first 15 minutes were about giving instructions for all participants including the rights of the participants, then 15 minutes for each group to watch their version of the sequences and fill out the questionnaire of the pilot test.

The next section explores in detail the statistical testing performed on the responses of the participants and analyzes the revealed results.

8.4/ Pilot Test Results

All the statistical tests performed in this section are based on Mann-Whitney U non-parametric tests, as we are evaluating, at a time, one ordinal dependent variable (5 responses of the participants to a question) based on one independent qualitative variable (2 groups of participants), and the sample size of all the groups $SampleSize < 30$. For all tests, the Confidence Interval CI is 95% so the alpha value $\alpha = 1 - CI = 0.05$, and the $p - value$ will be provided per test below.

Based on the responses of the participants of the pilot test organized in the groups N, D, and F, the comparisons of the responses, to investigate the role of explanations and the filtering of them, are presented in the following sections.

8.4.1/ No Explanation vs. Explanation

We compare the 11 participants of Group N (No explanation), on one hand, with the 16 participants that have received explanations of both Group D (Detailed explanation) and Group F (Filtered explanation) on the other hand.

Using a Mann-Whitney U test ($CI = 95\%, U = 45, p - value = 0.029$), Figure 8.2 shows the box plot that corresponds to the question: “Do you believe the only one time you watched the simulation tool working was enough to understand it?”, with 5 possible answers (*cf.* Section 7.3). The box plot shows that the median response of Groups D and F ($med = 4$) is significantly higher than the median response of Group N ($med = 2$), *i.e.* the participants that received explanations agree more than the participants with no explanation that watching the simulation once is enough.

Using a Mann-Whitney U test ($CI = 95\%, U = 43.5, p - value = 0.018$), Figure 8.3 shows the box plot that corresponds to the question: “How do you rate your understanding of how

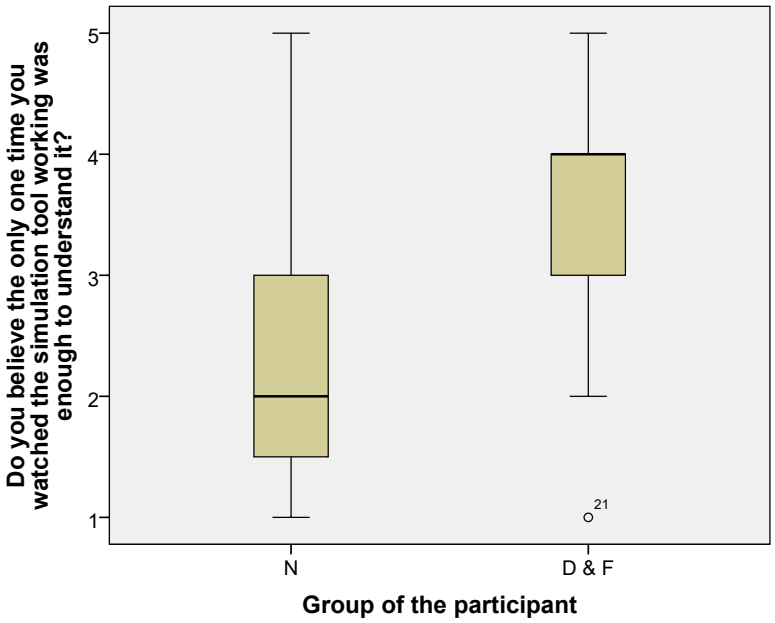


Figure 8.2: Pilot test: Do you believe the only one time you watched the simulation tool working was enough to understand it? (Explanation vs. No explanation)

the simulation tool works?", with the following possible answers: 5 (Very high), 4 (High), 3 (Neutral), 2 (Low), 1 (Very low). The box plot shows that the median response of Groups D and F ($med = 4$) is higher than the median response of Group N ($med = 3$), *i.e.* the participants that received explanations rate their understanding of the simulation with a higher value than the participants that did not receive any explanation.

According to these two results, RH1 (page 70) is proven, *i.e.* explainability increases the understandability of the human-on-the-loop in the context of remote agents. The respectful reader can notice that the questions of Figure 8.2 and Figure 8.3 have almost a similar goal. This is explained by the fact that when we have built the questionnaire, we have added some similar questions to assure the adherence and consistency of the responses of the participants.

8.4.2/ Detailed Explanation vs. Filtered Explanation

We compare the 8 participants of the Group D (Detailed explanations) on one hand with the 8 participants of the Group F (Filtered explanations) on the other hand.

Using a Mann-Whitney U test ($CI = 95\%, U = 15, p - value = 0.058$), Figure 8.4 shows the box plot that corresponds to the question: "Do you believe the only one time you watched the simulation tool working was enough to understand it?", with 5 possible answers (*cf.* Section 7.3). The box plot shows that the median response of Group D ($med = 4$) is

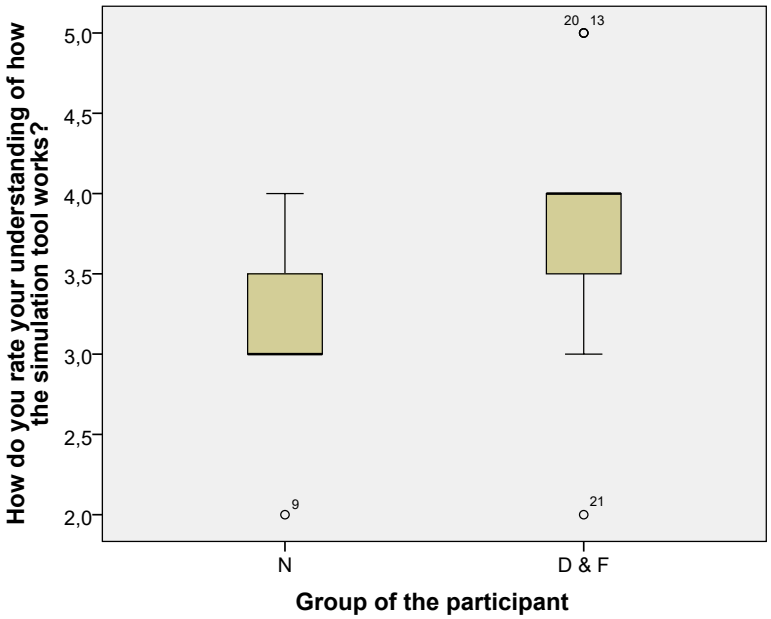


Figure 8.3: Pilot test: How do you rate your understanding of how the simulation tool works? (Explanation vs. No explanation)

higher than the median response of Group F ($med = 3$), *i.e.* the participants that received detailed explanations tend to agree more than those receiving filtered explanations that watching the simulation once is enough to understand it. This result could be explained by the fact that when a participant receives a lot of explanations, he/she tends to feel more confident that watching the simulation once is enough. However, it is worth mentioning here that the $p - value$ was slightly higher than the α value for this test, so this result is not decisively significant.

The last overwhelming sequence shown to the participants included 10 UAVs and 16 packages. For this sequence, we asked a specific question related to RH2. Using a Mann-Whitney U test ($CI = 95\%, U = 13, p - value = 0.044$), Figure 8.5 shows the box plot that corresponds to the question: "The explanation of how the simulation tool works in the last sequence has too many details", with 5 possible answers (*cf.* Section 7.3). The box plot shows that the median response of Group D ($med = 3.5$) is higher than the median response of Group F ($med = 2.5$), *i.e.* the participants that received detailed explanations were overwhelmed by the details of the explanations and think that the explanation was too much detailed compared to the participants that received filtered explanations.

Two findings can be deduced from the results of comparing Group D with Group F:

1. More details are preferable by the participant and it increases its confidence that watching the simulation once was enough to understand it, but with a questionable significance (Figure 8.4). This agrees with the findings of Harbers et al. [120] where

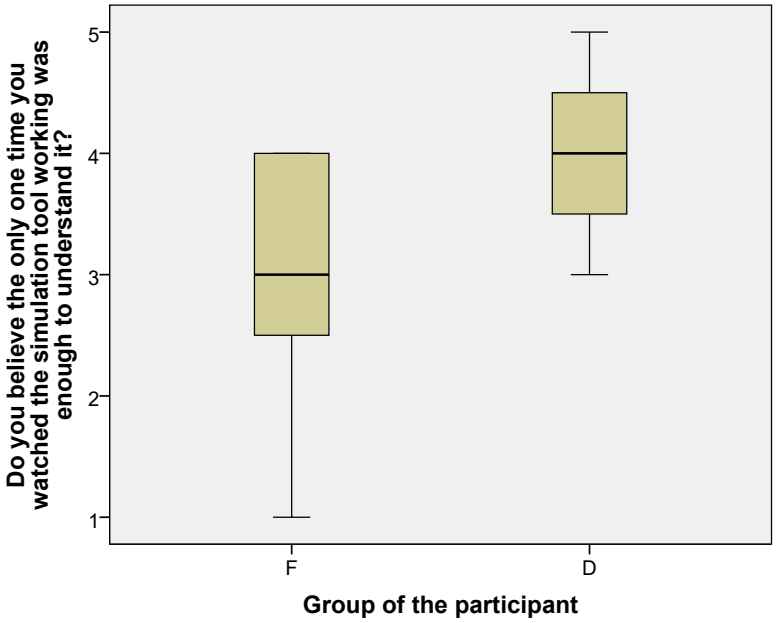


Figure 8.4: Pilot test: Do you believe the only one time you watched the simulation tool working was enough to understand it? (Detailed explanation vs. Filtered explanation)

it is mentioned that the participant prefers more details in the explanation.

2. However, with the increase of scalability, the participant is eventually overwhelmed with too many details (Figure 8.5) and in this case, the filtering of explanations is essential, and this proves RH2 (page 70). Moreover, the filtering of explanations gives more time for the participant to do other tasks and this aspect of shared autonomy could be investigated in future work.

8.5/ PILOT TEST LIMITATIONS

We tried to normalize the conditions of the test by providing the exact experimentation conditions for all participants. However, there may be still some personal factors that make the experience of each participant different. Additionally, when choosing a sample from the population, this sample may have traits that are not representative of the entire population (*e.g.* knowledge and interest in technology, culture, *etc.*) and that influence the responses of the questionnaire. Therefore, the generalization of the results is limited as the pilot test was conducted on a sample consisting of students (Bachelor, Master, and Ph.D.) in the technology and engineering domains which does not necessarily represent the whole population. However, there are many similarities between the participants in this test (*e.g.* age slice and major study).

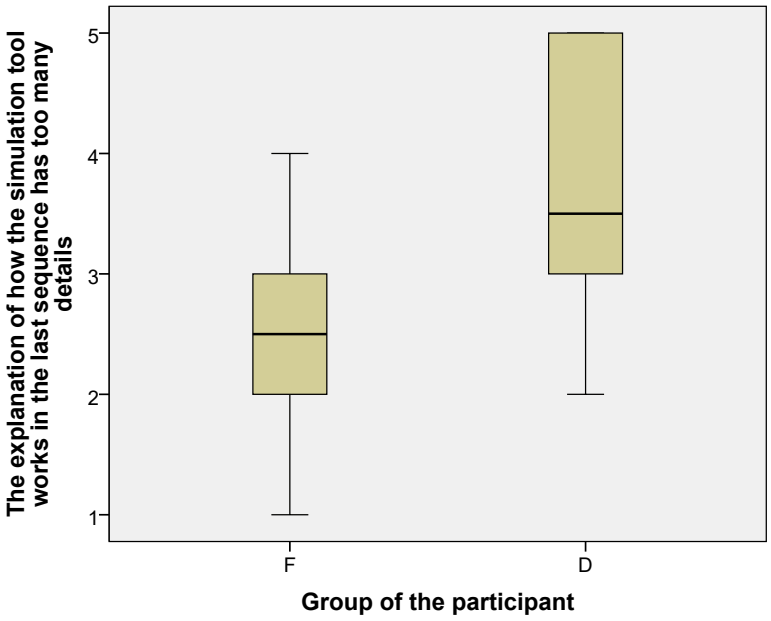


Figure 8.5: Pilot test: The explanation of how the simulation tool works in the last sequence has too many details (Detailed explanation vs. Filtered explanation)

8.6/ CONCLUSION

This chapter discussed the pilot test conducted to evaluate part of the contributions related to RQ1 (*see* Section 5.2.1) and in particular RH1 and RH2 (page 70). The experiment scenario in the pilot test is implemented using Repast Simphony [68] as an ABS framework that controls and manages the environment and the scheduler of the agents. The agents are all reactive agents and not BDI agents. The pilot test investigates the role of filtering of explanations provided by remote agents to the human. Three different cases are investigated: “No explanation”, “Detailed explanation” and “Filtered explanation”. Accordingly, the participants in the test have been organized into three groups: *(i)* Group N: the participants watch the simulation with no explanation; *(ii)* Group D: the participants watch the simulation with detailed explanations; *(iii)* Group F: the participants watch the simulation with filtered explanations.

The responses of the participants have been statistically analyzed, validated in terms of significance, and presented based on Mann-Whitney U non-parametric tests. According to the results of comparing the responses of Group N with Group D, explainability increases the understandability of the human-on-the-loop in the context of remote agents, *i.e.* RH1 is proven. Comparing the responses of Group D with Group F revealed that more details are preferable by the participant and it increases its confidence that watching the simulation once was enough to understand it but with a questionable significance. However, with too many details the participants are eventually overwhelmed and in this

case, the filtering of explanations is essential, *i.e.* RH2 is proven.

Mainly, and as a response to RQ1, we have aimed with the pilot test to reproduce the results of the literature regarding the benefits of explainability on one hand and the filtering of explanations on the other hand in the domain of remote robots (*e.g.* UAVs) represented as agents. The next chapter goes more steps forward to investigate various ways and manners to provide parsimonious explanations that strike a balance between simplicity and adequacy. It evaluates the full range of the contributions in the thesis.

9

MAIN TEST

9.1/ INTRODUCTION

In this chapter, we conduct the main empirical human study (or main test) based on the experiment application (*see* Chapter 7) of package delivery using civilian Unmanned Aerial Vehicles (UAVs) to answer two of the Research Questions (RQs) and evaluate two of the Research Hypotheses (RHs). In particular, the main test evaluates the part of the contribution related to RQ2: *Parsimonious Explanations* (Section 5.2.2) and RQ3: *Modeling Explainability for Remote Agents using Cognitive Architectures* (Section 5.2.3). In particular, the main test tackles and evaluates the following hypotheses (Section 5.3):

- RH3-1: **Adaptive filtering** with only normal explanations increases the understandability of the human-on-the-loop compared to **static filtering** with only normal explanations.
- RH3-2: **Adaptive filtering** with normal and **contrastive** explanations, *i.e. parsimony of explanations*, increases the understandability of the human-on-the-loop compared to **static filtering** with only normal explanations.
- RH3-3: **Adaptive filtering** with normal and **contrastive** explanations, *i.e. parsimony of explanations*, increases the understandability of the human-on-the-loop compared to **adaptive filtering** with only normal explanations.
- RH4: **Adaptive filtering** with normal and **contrastive** explanations, *i.e. parsimony of explanations*, increases the trust of the human-on-the-loop compared to **static filtering** with only normal explanations.

In this test, we investigate how parsimonious explanations could be formulated in Explainable Artificial Intelligence (XAI) by adapting the explanation phases (*see* Section 2.7):

- **Explanation Generation:** The process of explanation generation in the proposed architecture HAExA can be either *normal* or *contrastive*. Whereas the former is

sufficient for normal situations (*i.e.* situations where no unexpected events occur), the latter excels in explaining abnormal situations. For instance, the remote agent, here the UAV, could either generate raw normal explanations, *i.e.* non-contrastive explanations, like the one used in the pilot test (Chapter 8), or raw contrastive explanations based on the beliefs and intentions of the UAV.

- **Explanation Communication:** The way the raw explanation is communicated from the UAVs to the human is governed by the assistant agent. Two filters are investigated in this test: a static filter like the one used in the pilot test (Chapter 8), and an adaptive filter. The adaptive filtering is performed by the Belief–Desire–Intention (BDI) assistant agent whose beliefs and intentions change according to the complexity of the situation. Accordingly, it adapts the strictness of the filtering.

- **Combined Approach:** The explanations are formulated by combining aspects from both the generation and communication phases, *e.g.* combining adaptive filtering with contrastive explanations. In this approach, the formulation means that there is an intersection between the two phases, *i.e.* it is possible that the raw explanations are updated in the communication phase by the assistant agent, as in this later phase, the assistant agent has a general overview of the situation, and hence it could provide better context-aware explanations to the human.

In this test, the implementation includes BDI agents in the Agent-based Simulation (ABS) for realizing the human study to facilitate the reception of explanations (*see* Section 5.4.1). After watching the simulation and filling out the associated questionnaire, the responses undergo statistical testing to interpret, evaluate, and analyze the results (*see* Section 5.4.2). To validate the results, the significance of the responses of the participants is verified using both parametric and non-parametric tests.

The chapter is organized as follows. First, Section 9.2 outlines the methodology of conducting the main test. Second, Section 9.3 provides the statistical analysis and interpretation of the responses of the participants. Third, Section 9.4 outlines the limitations of the test. Finally, Section 9.5 concludes this chapter.

9.2/ MAIN TEST METHODOLOGY

This section conveys the methodology applied to conduct the main test as follows. Section 9.2.1 outlines the experimental details of the test. Section 9.2.2 details the process of organizing the participants in groups and how to aggregate their responses.

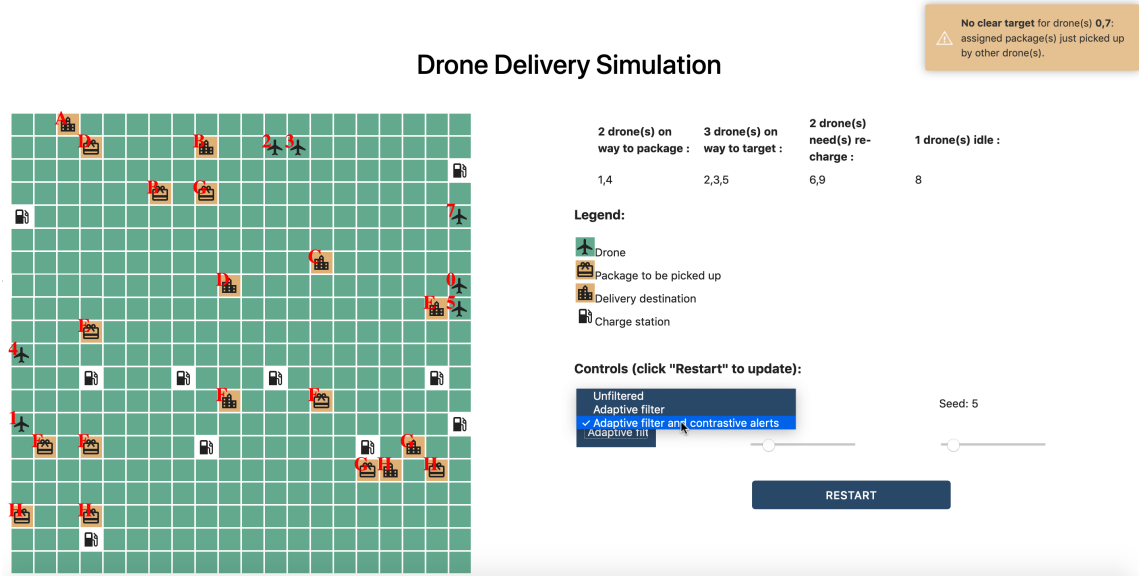


Figure 9.1: Main test simulation snapshot

9.2.1/ Experimental Details

The experiment scenario (*see* Section 7.2) in the main test is implemented using the JSson agent-oriented programming library [139, 206]. All agents in the simulation of the main test are BDI agents. The beliefs, desires, and intentions of these agents change according to the situation to formulate the parsimonious explanations that strike a balance between simplicity and adequacy. All the functionalities proposed in Chapter 6 are employed in this simulation. The simulation is run 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. Figure 9.1 depicts a snapshot of the simulation presented to the participants for the main test (*see* our work [206] for more technical details). 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.

In the case of the remote agents that represent the UAVs as an example of remote robots, the explanation formulation process helps in the explanation generation phase, to generate raw normal explanations in normal situations and raw contrastive explanations in abnormal ones based on the change of its beliefs and intentions, *i.e.* the remote agents are adaptive and context-aware when generating raw explanations. For the assistant agent, this helps in updating the raw explanations to ensure they have all the necessary information, *i.e.* adequacy, according to the combined approach between generation and communication phases. It also helps in filtering the updated explanations, *i.e.* simplicity, in overwhelming scenarios in the explanation communication phase based on the human cognitive load. Mainly, the assistant agent downgrades and upgrades the types of the explanations according to the levels of hierarchy illustrated in Figure 6.5 (page 99), *i.e.*

the assistant agent is adaptive and context-aware when communicating the explanations to the human (refer to Section 6.5.2 for more details).

9.2.2/ PARTICIPANTS AND GROUPS

The main test is conducted online, where the simulation is prepared as high-quality videos. The test instructions along with the links to the questionnaire and the videos are provided in a presentation provided in a signal link. To reach the participants of the test, we have broadcast the link of the main test in mailing lists in which we are affiliated. Moreover, we have also posted the link of this test to social networks. To obtain the sample used in the analysis of the main test, we focus on voluntary sampling. The people who receive the link choose to participate or not in our experiment. Voluntary sampling has some advantages such as the simple way to conduct the test, inexpensiveness, easy data collection, easy access, *etc*. However, it has also some drawbacks such as response biases, *i.e.* sample members are self-selected volunteers. Therefore, and unlike in the pilot test, the participants in the main test are not restricted to university students only, but the sample includes a broader audience. However, only the participants with *B1* English level or more can participate as all the explanations are communicated in English. Voluntarily participants watch the simulation and then fill out a questionnaire (*see* Appendix D).

The representative sample is composed of 90 participants. They were randomly divided into three groups (*SF*, *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 only and static filtering;
2. Group *AF* (30 participants) watches the simulation with normal explanations only and adaptive filtering;
3. Group *AC* (30 participants) watches the simulation with normal and contrastive explanations and adaptive filtering.

After watching the simulation sequences assigned to the participants, they fill out the questionnaire of the test. Apart from the first 8 questions of Participant Details, Functionalities categories (refer to Section 7.3.1 for more details on the categories of the questions) and excluding the question Q18 about curiosity, the questions understudy in the statistical analysis of this test are 12 questions. (refer to Appendix D for all the questions of the questionnaire in the main test). The responses are distributed to a 5-points Likert scale [132] (refer to Section 7.3 for more details on building the questionnaire). These 12 questions are analyzed and discussed in Section 9.3 (refer to Table 9.1 for the list of these questions).

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 $\overline{x}_{age} = 26.44$, and standard deviation of age $s_{age} = 7.348$). Regarding previous knowledge of the participants about UAVs, they have self-rated their knowledge using 5-points Likert as (mean of UAV knowledge $\overline{x}_{UAV\_knowledge} = 3.27$, and standard deviation of UAV knowledge $s_{UAV\_knowledge} = 1.1$). Therefore, it can be noticed that the randomly selected participants of the test are heterogeneous regarding their age, sex, and knowledge of UAVs.

To validate the results, the significance of the responses of the participants has been verified using statistical testing. To avoid biases in the data analysis and due to the dispute between researchers and statisticians (*see* Section 5.4.2), the methodology adopted in the main test is to conduct both the parametric test that is ANOVA and the non-parametric test that is Kruskal-Wallis for the data analysis (*see* Section 5.4.2). The next section explores in detail the statistical testing performed on the responses of the participants and analyzes the revealed results.

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

Table 9.1: The $p-value$ of each investigated question in the main test for both ANOVA and Kruskal-Wallis test

9.3/ MAIN TEST RESULTS

In terms of parametric tests, the homogeneity of variance of our data, which is also needed for the use of parametric testing, is verified. On contrary, if we focus on non-parametric tests such as Kruskal-Wallis test for the data analysis, we are also faced with the well-known problem for the data analysis of the ordinal data. To use Kruskall-Wallis, the data should not have too many ex-aequo, However, in our data, we have only five different categories of data, and all the responses of the 90 participants are spread, distributed, and divided among these five categories, *i.e.* there are many ex-aequo.

This section is organized as follows. First, Section 9.3.1 verifies the initial significance of the results. Second, Section 9.3.2 thoroughly analyzes and interprets the results to evaluate the RHs. Finally, Section 9.4 outlines the main test limitations.

9.3.1/ INITIAL VERIFYING OF THE SIGNIFICANCE

For each of the 12 questions understudy, the null hypothesis is $H_0 : \mu_{SF} = \mu_{AF} = \mu_{AC}$ for ANOVA (respectively 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).

Table 9.1 outlines the statistically significant results obtained by both ANOVA and Kruskal-Wallis (KW) in our main test. As presented by this table, although the $p-values$ of ANOVA and Kruskal-Wallis are different, the results of significance are similar between these two tests, *i.e.* when the null hypothesis is rejected by ANOVA, it is also rejected by Kruskal-Wallis, and the same is true for acceptance. Since we got the same significance results for both ANOVA and Kruskal-Wallis and due to the power of parametric tests, *i.e.* they give better results to reject the null hypothesis, as according to Dr. Geoff Normann, parametric tests are generally more robust than non-parametric tests [269], the rest of the data analysis is done with ANOVA. In particular, for the pair-wise comparison with ANOVA, this test focuses on Tukey Honest Significant Difference (Tukey HSD) test because all the groups have the same size (30 participants per group), and the homogeneity of variance is verified by the data. To this end, Table 9.2 outlines the comparison results obtained by Tukey pair-wise ANOVA for all the questions that provided significant $p-values$ in Table 9.1. In Table 9.2, we recall that *AF* means adaptive filtering with normal explanations only, *AC* means adaptive filtering with normal and contrastive explanations, and *SF* means static filtering with normal explanations only.

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

Table 9.2: Tukey HSD pair-wise ANOVA comparisons of the groups in the main test

9.3.2/ Data Analysis with Parametric Testing

9.3.2.1/ Investigating the Understandability

The questions Q9, Q10, Q11, Q12, Q19, Q20, and Q21 are considered. All the obtained $p-values$ of these questions, except for Q20[1], are significant, *i.e.* the $p-values$ obtained by both ANOVA and Kruskal-Wallis tests outlined in Table 9.1 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, and Q21 are in bold font in Table 9.2. They have significant comparable results discussed between groups in pairs as follows:

- AF vs. SF pairwise comparison: The results (Table 9.2) reveal that the questions Q11, Q19, Q21 (*see* box plots in figures 9.4, 9.6, and 9.7 to visualize the responses of participants) 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 (*see* box plots in figures 9.2, 9.3, and 9.5 to visualize the responses of participants), the differences between the means of *AF* and *SF* are not statistically significant in Table 9.2 ($p-value > .05$). 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 RH3-1 (adaptive filtering increases the understandability compared to static filtering) for all the six questions.

  In the pilot test, it was proven 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 deceive added value in increasing the understandability. This result may be explained by the fact that for abnormal situations, the participants did not understand well the situation. Therefore, the hypothesized solu-

[1] *see* Appendix E for the box plots of the insignificant results of this question.

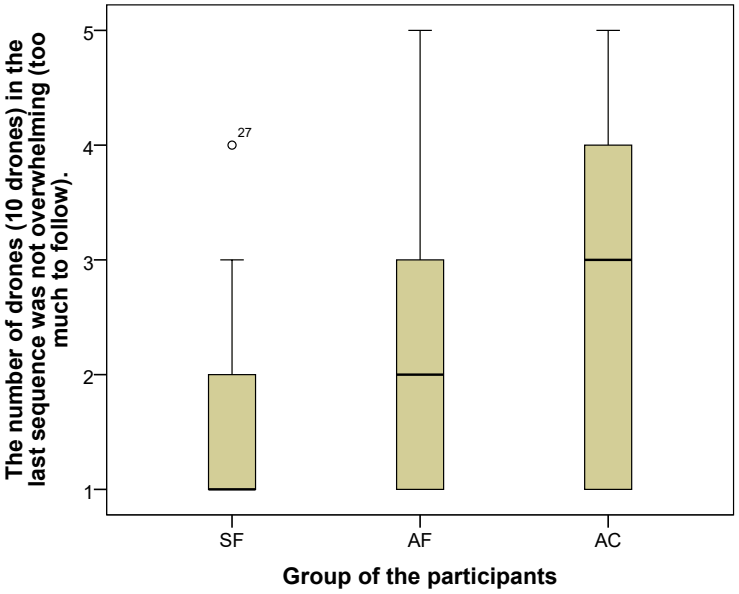


Figure 9.2: Main test: Q9 ($\overline{x}_{SF} = 1.67, \overline{x}_{AF} = 2.33, \overline{x}_{AC} = 2.63$, medians are represented in the figure)

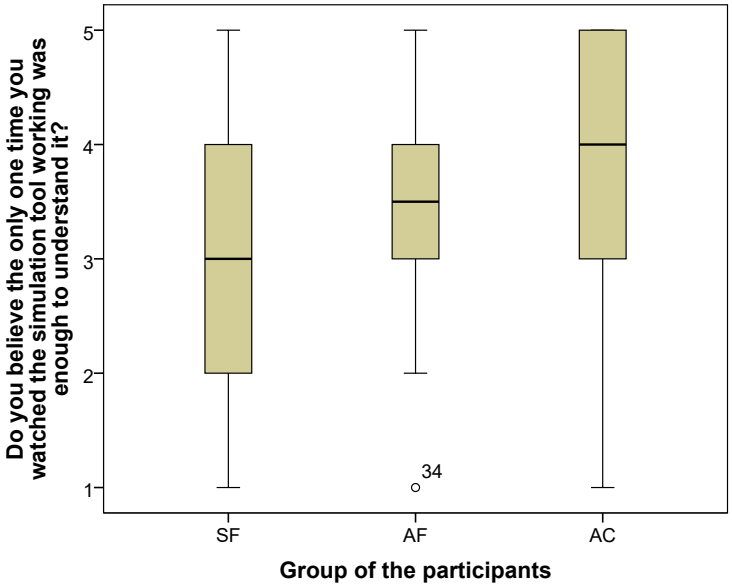


Figure 9.3: Main test: Q10 ($\overline{x}_{SF} = 2.87, \overline{x}_{AF} = 3.47, \overline{x}_{AC} = 4.00$, medians are represented in the figure)

tion in the explanation formulation in HAExA was to handle the abnormal situations using contrastive explanations in terms of explanation generation (the case of *AC*), *i.e.* an explanation formulation as a combination of explanation generation and communication.

- AC vs. SF pairwise comparison: The results (Table 9.2) show that all the results

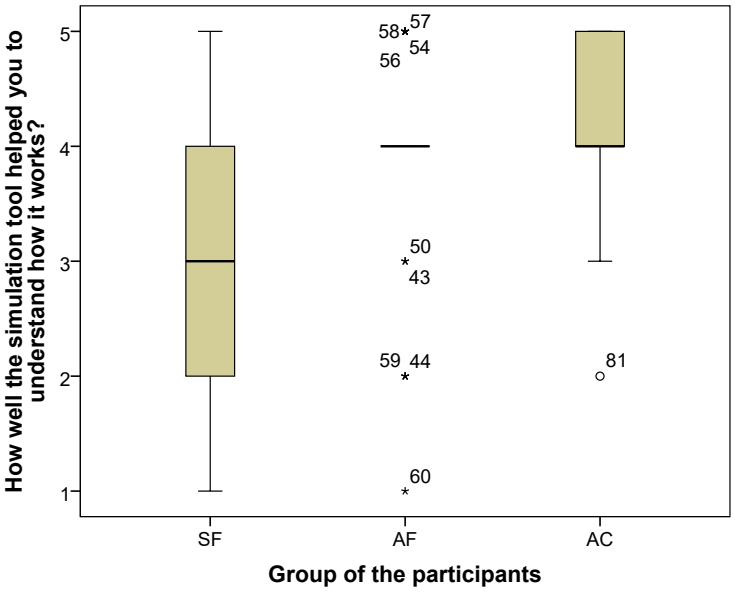


Figure 9.4: Main test: Q11 ($\overline{x}_{SF} = 3.13, \overline{x}_{AF} = 3.87, \overline{x}_{AC} = 4.33$, medians are represented in the figure)

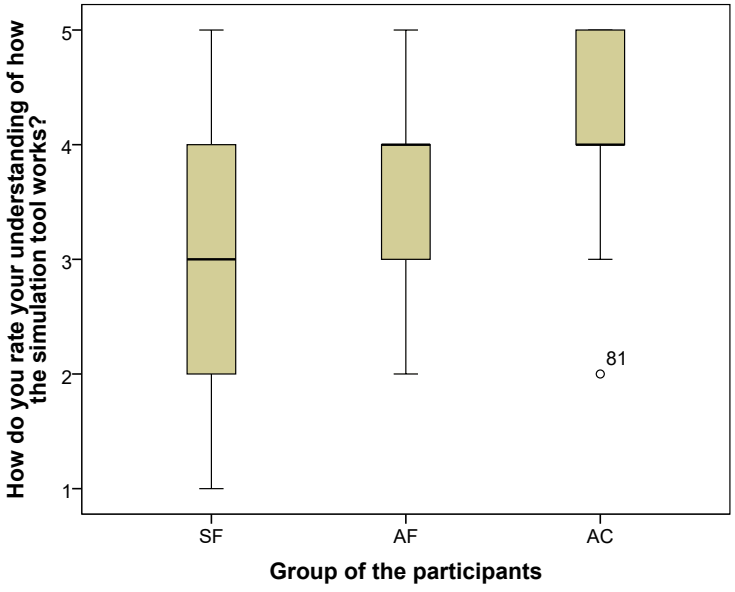


Figure 9.5: Main test: Q12 ($\overline{x}_{SF} = 3.23, \overline{x}_{AF} = 3.77, \overline{x}_{AC} = 4.23$, medians are represented in the figure)

for the questions understudy Q9, Q10, Q11, Q12, Q19, and Q21 (refer to box plots in figures 9.2, 9.3, 9.4, 9.5, 9.6, and 9.7 to visualize the responses of participants) have significant differences between the means of *AC* and *SF* ($p - value \leq .05$), *i.e.* we can reject the null hypothesis $H0$ and conclude that the means of *AC* and *SF* are not equal. For all these six questions, the mean differences are positive in the

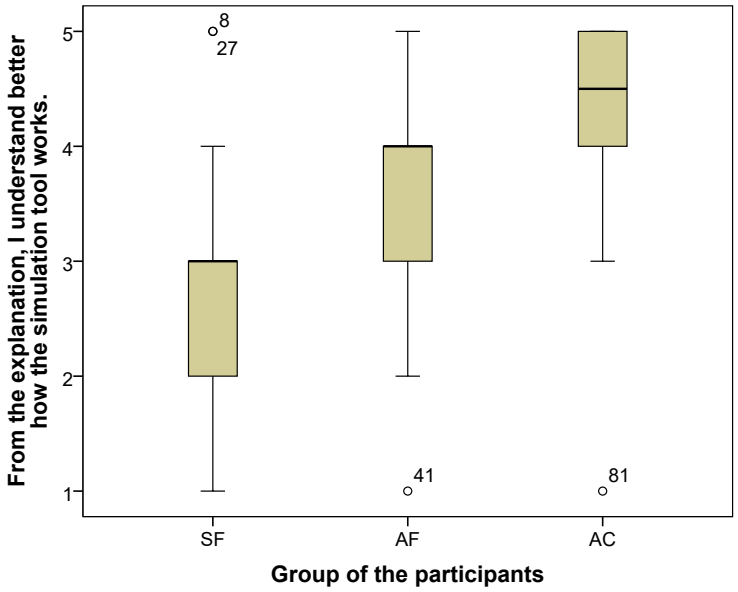


Figure 9.6: Main test: Q19 ($\overline{x}_{SF} = 2.83, \overline{x}_{AF} = 3.70, \overline{x}_{AC} = 4.20$, medians are represented in the figure)

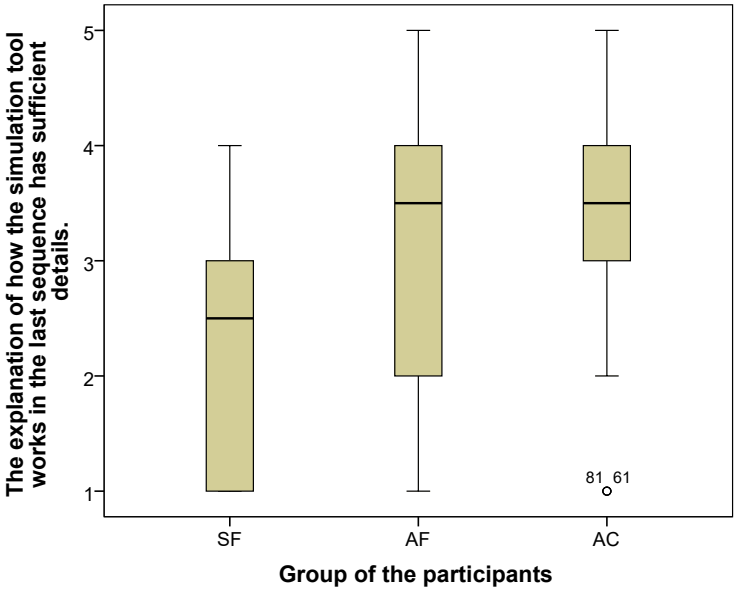


Figure 9.7: Main test: Q21 ($\overline{x}_{SF} = 2.57, \overline{x}_{AF} = 3.53, \overline{x}_{AC} = 3.57$, medians are represented in the figure)

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 both normal and contrastive explanations with adaptive filtering (*AC*) agree more that this explanation formulation

is more understandable than the one that includes only normal explanations with static filtering (*SF*). In other words, the results show that empowering HAExA with contrastive explanations in the generation phase followed by updating them in the communication phase and adaptive filtering in the communication phase provides the necessary concise information, *i.e.* parsimonious explanations, for the human to better understand the situation. This means the research hypothesis RH3-2 (page 70) is accepted.

- AC vs. AF pairwise comparison: For all the questions understudy 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 RH3-3 (page 70) is rejected.

  It is worth mentioning here, that even though the results of *AC* are not significantly better than those of *AF*, this does not mean that the results of *AF* are significantly better than those of *AC*. It just means that we cannot confirm if there is a difference between the two groups, and further research investigations with larger sample sizes could produce better results. This result opens the door for the trade-off that emerges when using contrastive explanations where the benefit could be situational in abnormal situations. A new research hypothesis could be tested in the future: The more the abnormal situations, the more the contrastivity in generating the explanation is needed. To provide a parsimonious explanation, it is vital to pinpoint the unnecessary information, and for that, considering the parts of a contrastive explanation to be necessary or not is a future challenge.

The results in general show that *AC* is firmly better than *SF*, while *AF* being better than *SF* is questionable. However, the direct comparison between *AC* and *AF* shows no significant difference between them to say which is better. Even though *AC* is not decisively better than *AF* in a pairwise 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 (normal and contrastive explanations) and explanation communication (adaptive filtering), 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 with only normal explanations, *i.e.* without contrastive explanations, is not necessarily better than static filtering with only normal explanations in

all situations. This means that adapting only the communication phase of the explanation is not enough to increase the understandability of humans of the explanations, as there is a need for adapting also the generation phase of the explanation. In other words, going one step forward by relying on adaptive filtering, *i.e.* achieving only simplicity, is not enough to provide parsimonious explanations.

- Adaptive filtering empowered by normal and contrastive explanations, *i.e.* parsimony of explanations, is better than static filtering with only normal explanations in increasing the understandability of humans of the explanations.
- Even though it is situational to consider that contrastive explanations are better than normal explanations, they can be used in all cases with no fear of overwhelming the human.
- Having only a context-aware generation of explanations based on the beliefs and the intentions of the remote agents is not enough. There is a need for updating the raw explanations by the assistant agent that has a global view of the context. In other words, the adaptation in terms of the generation of explanations should not be only in the generation phase, instead, it should be a combined effort in both the generation and communication phases.
- The benefits of adopting an agent-based approach are twofold. First, it helps in realizing the explanation formulation process, as it organizes the various interactions between the entities in the system. Second, ABS provides a test-bed environment to conduct human studies and enriches the XAI domain by allowing for more empirical and results-oriented research that facilitates the explanation reception by the human. Additionally, ABS represents a good means to visualize the behavior of the remote robots, represented as agents, that are not co-located with the human, hence better explanation reception.

Therefore, the results above confirm our proposal that the best explanation formulation integrates the phases of explanation generation and communication in a context-aware and adaptive combination thereby striking a balance between simplicity and adequacy. Additionally ABS has been used in this test as a good tool to facilitate the last phase of explanation reception. However, future work should be done to integrate this phase in the proposal.

9.3.2.2/ INVESTIGATING THE TRUST

The questions Q13, Q14, Q15, Q16, and Q17 are considered regarding RH4 (page 70) that investigates the trust of the participants regarding the explanation. The obtained

$p - values$ of all the questions Q13, Q15, Q16, and Q17 are not statistically significant (*see* Table 9.1). The only question with a significant $p - value$ is question Q14. However, Q14 has no significant value in the pairwise comparison between the groups (*see* Table 9.2), so it is discarded[2]. Therefore, we cannot reject the null hypothesis $H0$ and we reject RH4.

This result confirms previous results found in a similar context in the literature [185], and a related work when building human users' mental models of how an agent works [162]. The literature of virtual agents and social agents and robots may help in the direction of increasing the trust. Moreover, more work should be done to promote trust as the participants do not yet trust the remote agents even with explanations.

9.4/ Main Test Limitations

As stated before, participants involved in this test watched the simulation online and filled out the questionnaire online. Therefore, some limitations can be considered:

- **Sampling bias:** Although we have tried to broadcast the requests of participation of this test as much as possible on the Internet, some voluntary participants are close to our networks. Certain categories or age groups remain difficult to reach via the Internet, and therefore, the participants could not represent the entire heterogeneous population.

- **Lack of contact:** Because the participants filled out their questionnaires online, they may have not understood some questions and we cannot guarantee that all the participants have well understood all the questions although precision and conciseness were considered when building the questionnaire.

- **Different technology infrastructure:** Participants could use different hardware such as smartphones, laptops, desktops, *etc*. to participate in the test. This means having different properties to fill out the questionnaire, with different web browsers on different operating systems, various Internet speed, *etc*. Additionally, the participants could not have the same conditions when conducting the test, *i.e.* same place, time, quality of presenting medium, *etc*. This could affect the reception of the explanations by the participants.

[2] *see* Appendix E for the box plots of the insignificant results of the questions Q13, Q14, Q15, Q16, and Q17.

9.5/ CONCLUSION

This chapter discussed the main test conducted to evaluate part of the contributions related to RQ2 and RQ3 and in particular RH3 and RH4 (page 70). The experiment scenario in the main test is implemented using the JS-son agent-oriented programming library where all the agents were represented as BDI agents. The main test investigates the role of the explanation formulation process that allows providing parsimonious explanations by remote agents to the human. Three different cases were investigated where the participants in the test have been organized into three groups: *(i)* Group *SF* watches the simulation with normal explanations only and static filtering; *(ii)* Group *AF* watches the simulation with normal explanations only and adaptive filtering; *(iii)* Group *AC* watches the simulation with normal and contrastive explanations and adaptive filtering.

The responses of the participants have been statistically analyzed, validated in terms of significance, and presented based on Kruskal-Wallis non-parametric tests and ANOVA parametric tests.

The test investigates two points: *(i)* Understandability: It is proved that there is a need to have a combination of the phases of the explanation generation and communication to formulate the most useful explanation for the human. Additionally, this formulation is context-aware, *i.e.* specific levels of these two phases are used according to the context. Comparing several combinations of explanation formulation, it is proven that the best one includes using adaptive filtering with both normal and contrastive explanations, *i.e.* RH3-1 and RH3-3 are rejected while RH3-2 is accepted. *(ii)* Trust: no combination managed to increase significantly the trust of the human, *i.e.* RH4 is rejected.

In summary, and as a response to RQ2, to achieve the parsimony of explanations, the explanation formulation process must be context-aware and adaptive to integrate both the generation and communication phases of explanations. As a response to RQ3, the results obtained by HAExA suggest that the benefits of adopting an agent-based architecture, and in particular a cognitive BDI architecture, are twofold. First, it helps in realizing the explanation formulation process, as it organizes the various interactions between the entities in the system and allows for an adaptive and context-aware response based on the changes in the beliefs and desires of the agents, *i.e.* a similar way to how humans cognitively handle the explanations. Second, ABS provides a test-bed environment to conduct human studies that facilitate the reception of the explanations by the human and enriches the XAI domain by allowing for more empirical and results-oriented research. The next chapter concludes the thesis and highlights future perspectives.

V

Conclusion and Perspectives

10

General Conclusion

10.1/ Summary of the Ph.D. Thesis

In the future Artificial Intelligence (AI) systems, it is vital to guarantee a smooth human-agent interaction, as it is not straightforward for humans to understand the agent's state of mind, and explainability is an indispensable ingredient for such interaction. Accordingly, the research domain of Explainable Artificial Intelligence (XAI) is gaining increased attention from researchers of various disciplines.

When providing explanations to humans, the aim is to imitate how humans generate and communicate explanations in their everyday life. Everyday explanations are the explanations of why particular events, behaviors, decisions, *etc*. happened. Investigating everyday explanations leads us to discuss the parsimony of explanations that could help in providing the necessary information while reducing the human cognitive load to avoid overwhelming the human with useless information, *i.e.* to achieve the parsimony of explanations, there is a trade-off between the two features of an explanation, namely simplicity and adequacy. In Chapter 3, we have outlined how works in the literature have started to respond to these two features. Some works tackled the simplicity of explanations [121, 122, 120, 160], and others investigated the adequacy [145, 60, 202, 241, 302, 214, 56, 268], while few works tried to handle the trade-off between simplicity and adequacy [162]. Thus, we have identified the related open research issues to be tackled.

The problem of understanding the behavior of robots is more accentuated in the case of remote robots since —as confirmed by recent studies in the literature [124, 23]— remote robots, *e.g.* Unmanned Aerial Vehicles (UAVs), tend to instill less trust than robots that are co-located. In this context, an obvious research frontier for the autonomous agents and Multi-agent Systems (MAS) community is the design of explainable intelligent agents [207]. This thesis is a part of the academic project UrbanFly and one of the goals of this project is to propose novel models for simulating UAVs in urban environments and

smart cities. In this context, UAVs represent the remote robots explaining their environment, behavior, and actions to the human. Accordingly, in Chapter 4, we have conducted a Systematic Literature Review (SLR) of Agent-based Simulation (ABS) for UAVs.

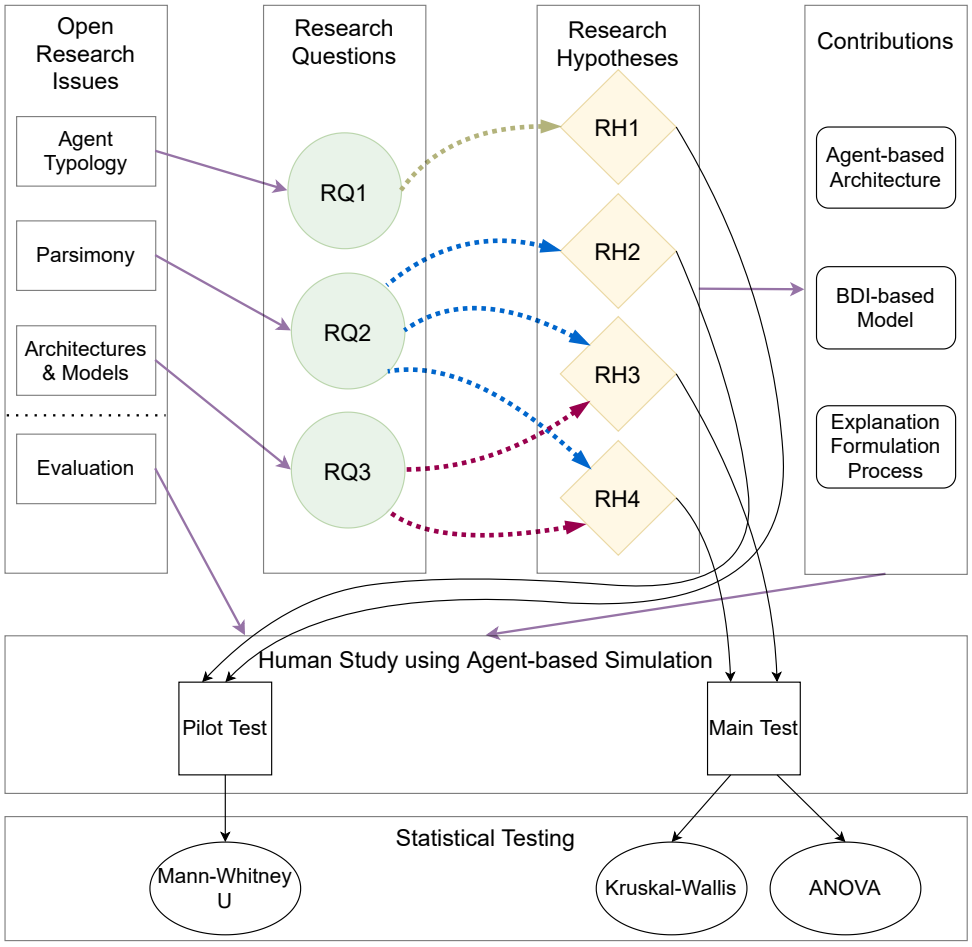


Figure 10.1: Research methodology of the thesis (version-2)

The research methodology conducted in the thesis is five-fold (Figure 10.1):

(i) Identify open research issues after analyzing the related work.

(ii) Define the *Research Questions (RQs)* based on the identified research issues.

(iii) Structure the RQs in *Research Hypotheses (RHs)* that can be statistically analyzed.

(iv) Propose the architecture, the model, and the process to answer the RQs.

(v) Conduct a specific experimental methodology to evaluate the proposals by statistically investigating the RHs according to the recommendations in the XAI domain.

Chapter 6 has explored in detail the contributions of the thesis. We have proposed an agent-based explainability architecture, named HAExA, to facilitate the human-agent ex-

plainability considering that agents represent remote robots. The main goal of the contributions is to provide parsimonious explanations to the human that strike a balance between adequacy and simplicity. A model based on Belief–Desire–Intention (BDI) has been proposed in HAExA to represent all agents. In principle, HAExA includes two types of agents: *(i)* The remote agents as part of the environment. *(ii)* The assistant agent whose role is to be an interface between the remote agents and the human.

HAExA investigates the three phases of providing an explanation from agents to the human: generation, communication, and reception (*see* Section 2.7). We argue that a well-formed *adaptive* and *context-aware* combination of these phases leads to formulating a parsimonious explanation. This process seeks to maximize the explanation's adequacy concerning an AI system while minimizing its impact on the human's cognitive load. To achieve this, the proposed architecture relies on an *explanation formulation process*. Section 6.5 proposed and thoroughly explored the explanation formulation process. This section was divided into two main subsections. First, Section 6.5.1 covered the generating of *raw explanations* by the remote agents. They have two main types: *Normal* in relatively normal situations, and *contrastive* in abnormal ones. These remote agents are BDI agents, whose beliefs and intentions are used to generate raw explanations. Section 6.5.2 tackled the communication phase of providing an explanation. The assistant agent has a global view of the context thanks to the raw explanations and messages it receives from the remote agents. Accordingly, it adaptively *updates* the raw explanations based on the changes in its beliefs and intentions, *i.e.* the assistant agent is context-aware. It mainly *downgrades* or *upgrades* the raw explanations according to a defined *hierarchy of explanations* that considers the trade-off between adequacy and simplicity. Additionally, it filters the updated explanations respecting thresholds of the human cognitive load.

The human understandability of AI is subjective, and this emphasizes the importance of empirical human studies where the opinions of humans on the usefulness of explanations are collected and analyzed. In this thesis, we design and conduct empirical human-agent interaction studies with the help of human participants to evaluate the proposed architecture and to investigate the RHs. These studies rely on well-established XAI metrics and questionnaires (*see* Section 2.10) in the literature.The studies are based on an application of package delivery using civilian UAVs, as an example of remote robots represented as agents, and implemented using ABS tools. Chapter 7 presented this application with the experiment scenario. We have conducted two tests based on the experiment scenario: The pilot test (Chapter 8) and the main test (Chapter 9).

Chapter 8 has discussed the pilot test conducted to evaluate part of the contributions related to RQ1 (*see* Section 5.2.1) and in particular RH1 and RH2 (page 70). The experiment scenario in the pilot test was implemented using Repast Simphony [68] where the agents are all reactive agents. The pilot test investigates the role of filtering of expla-

nations provided by remote agents to the human. The responses of the participants are statistically analyzed, validated in terms of significance, and presented based on Mann-Whitney U non-parametric tests. The results have shown that the explanation increases the ability of human users to understand the simulation. but too many details overwhelm them; then, the filtering of explanations is preferable. Mainly, we aimed with the pilot test to reproduce the results of the literature regarding the benefits of explainability on one hand and the filtering of explanations on the other hand in the domain of remote robots (*e.g.* UAVs) represented as agents.

In chapter 9, the main test evaluated the part of the contributions related to RQ2 (*see* Section 5.2.2) and RQ3 (*see* Section 5.2.3) and in particular RH3 and RH4 (page 70). The main test investigates the role of the explanation formulation process that allows providing parsimonious explanations by remote agents to the human. In the main test, we have conducted both a *Kruskal-Wallis* non-parametric test and the *ANOVA* parametric one to evaluate, analyze, and validate the results. The test investigates two points: *(i)* Understandability: It is proved that there is a need to have a combination of the phases of the explanation generation and communication to formulate the most useful explanation for the human. Additionally, this formulation is context-aware, *i.e.* specific levels of these two phases are used according to the context. Comparing several combinations of explanation formulation, it is proven that the best one includes using adaptive filtering with both normal and contrastive explanations, *i.e.* RH3-1 and RH3-3 are rejected while RH3-2 is accepted. *(ii)* Trust: no combination managed to increase significantly the trust of the human, *i.e.* RH4 is rejected.

In summary, the thesis responses to the posed RQs as follows:

- RQ1. Does explainability increase the humans' understandability of the remote robots represented as agents? On the one hand, and regarding understandability, explainability increases the understandability of the humans in the domain of remote robots (*e.g.* UAVs) represented as agents. However, with overwhelming situations, the filtering of explanations that provides less, concise, and synthetic explanations is needed to adhere to the human cognitive load. On the other hand, it is not decisive to confirm the same findings regarding the trust.
- RQ2. How to strike a balance between simplicity and adequacy? The parsimony of explanations is effective to handle the trade-off between simplicity and adequacy. To achieve the parsimony of explanations, the explanation formulation process must be context-aware and adaptive to integrate both the generation and communication phases of explanations. The results have revealed that adapting only the communication phase of the explanation, via the filtering of explanations, is not enough to increase the understandability, as there is a need for adapting also the generation phase of the explanation. Moreover, having only a context-aware generation

of explanations based on the beliefs and the intentions of the remote agents is not enough. There is a need for updating the raw explanations by the assistant agent that has a global view of the context. In other words, the adaptation in terms of the generation of explanations should not be only in the generation phase, instead, it should be a combined effort in both the generation and communication phases. In summary and for the contributions to give the full potential, both the combined generation and updating of normal and contrastive explanations on the one hand and the adaptive filtering of explanations on the other hand are needed.

- RQ3. Are the cognitive architecture and the BDI model good candidates for human-agent explainability? The results obtained by HAExA suggest that the benefits of adopting an agent-based architecture, and in particular a cognitive BDI architecture, are twofold. First, it helps in realizing the explanation formulation process, as it organizes the various interactions between the entities in the system and allows for an adaptive and context-aware response based on the changes in the beliefs and intentions of the agents, *i.e.* a similar way to how humans cognitively handle the explanations. Second, ABS offers a test-bed environment to conduct human studies that facilitate the explanations reception by the human and enriches the XAI domain by allowing for more empirical and results-oriented research. Additionally, ABS represents a good means to visualize the behavior of the remote robots, represented as agents, that are not co-located with the human.

The rest of this chapter explores our vision of future aerial transport systems in smart cities. Then, it discusses the derived research directions in terms of explainability and the domain of UAVs.

10.2/ PERSPECTIVES

This section outlines our future perspectives related to the thesis and is organized as follows. Section 10.2.1 conveys our vision of the future aerial transport systems in smart cities where explainability influences significantly the services offered to citizens. Section 10.2.2 highlights the research directions in the domain of explainability based on our vision of the future smart cities, while Section 10.2.3 highlights the directions related to UAVs.

10.2.1/ FUTURE AERIAL TRANSPORT SYSTEMS IN SMART CITIES

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

demands of sustainable urban environments and improve the quality of life for citizens, municipalities will increasingly rely on novel transport solutions. In particular, UAVs are expected to have a crucial role in the future smart cities thanks to their interesting features such as autonomy, flexibility, mobility, adaptive altitude, and small dimensions. However, densely populated megalopolises of the future are administrated by several municipals, governmental and civil society actors, where vivid economic activities involving a multitude of individual stakeholders take place. In such megalopolises, the use of agents for UAVs is gaining more interest especially in complex application scenarios where coordination and cooperation are necessary. This section sketches a visionary view of the UAVs' role in the transport domain of future smart cities.

While other work [203, 195] considered the challenges of deploying UAVs in smart cities, they mainly mentioned the existing research with no focus on the role of agents. Consequently, the vision exposed in this section focuses on the application of multiagent concepts to smart cities, and specifically to UAVs in these cities. Several research directions related to multiagent systems in this context are proposed.

Today, the need for transport of people and goods is increasing but so is traffic congestion, air pollution, road accidents, and climate change. Some of the solutions for these problems come in a form of ride-sharing [12, 257]. However, in the future, cities will need to rely on "high-tech" mobility solutions including the Internet of Things (IoT) and UAV technologies. Thanks to their autonomy, flexibility, mobility, low-cost maintenance, and coverage, UAVs are a useful solution for many of the transport challenges. Moreover, the need for explainability is evident in such solutions considering that the main clients of such solutions are humans. The explanations are provided through specific private means (7) or public ones (8). Below are listed our key visions of the future of transport with UAVs in smart cities where explainability plays a vital role (refer to our work [208] for more details):

- **In-the-Air Services:** Recent applications of UAVs concern the fast delivery of goods, such as commercial products ((1) in Figure 10.2), medical products or first aid kit (2), or food (3). In the future, other entities may be carried by UAVs, such as passengers (4) and big/heavy containers (5). These entities could be carried by a single UAV, or by a swarm of UAVs (6).
- **Smart transport and Traffic Management:** Another vital contribution of UAVs is smart transport, which will likely be another key area of development for any future smart city. Every city can rely on UAVs for improving urban transport and creating a sustainable ecosystem. For example, a flying UAV can explain and guide pedestrians on the ground via smart devices (7) or explanation panels (8), or guide other UAVs in the air through Vehicular Ad-hoc Networks (10). The latter are known as smart traffic management UAVs that control the flow of people and goods in the

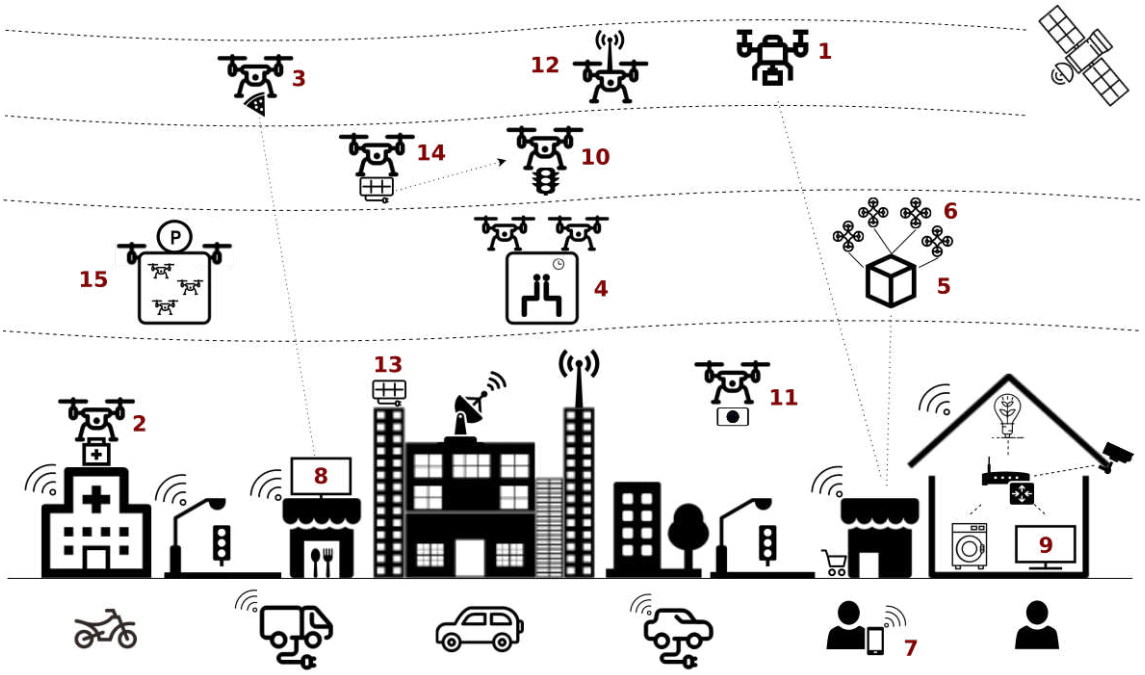


Figure 10.2: UAVs in a future smart city

sky. Air traffic management will divide the sky into free-flying areas and corridors in which all the UAVs will fly in the same direction. For example, heavy lifting (5) and passenger transport (4) can occupy the lowest corridor.

- **Air and Climate Management:** As cities inevitably become busier, the quality of the air, climate and, noise levels created by city systems can be monitored (11) and citizens can be informed through explanations (8 or 9). Municipalities can act with real-time actions to better manage the comfort and health of the citizens. To make the transport infrastructure of the future more sustainable, zero-emission and low-noise electric power is the solution. Using electric vehicles like UAVs tends to offer a silent, clean, emission-free, and resource-efficient city minimizing the risks affecting the health and safety of citizens. Furthermore, UAVs can be equipped with sensors. Data gathered from these sensors can be used by stakeholders to build a map of the environment state, such as air pollution and noise. Yet, a huge number of flying UAVs in the sky raise some environmental issues such as the recycling of out-of-service UAVs and the supply of clean and renewable energy to charge them.

- **In-the-Air Infrastructures:** Connecting objects within the smart cities via wireless technologies is already a reality (*e.g.* WiFi, 4G/5G, satellite, *etc.*). UAVs can offer a novel communication infrastructure by providing communication nodes where static/conventional nodes cannot be present (12). Because of the UAV mobility, this infrastructure may be deployed dynamically, even when the ground communication infrastructure cannot be used, *e.g.* in case of a natural disaster (see below). An-

other example is when the population density increases at a specific location for a limited time, *e.g.* at a football stadium. In this case, UAVs can offer an efficient networking service to the spectators. Energy consumption of UAVs is an issue: the average flying time for civil multirotor UAVs is around 20 minutes [191]. Consequently, it is mandatory to supply energy charging services to the UAVs on the buildings (13). This service may be also provided on-the-fly by other UAVs with eco-friendly solar power systems (14). Other types of infrastructures may appear in future smart cities, such as aerial parking areas (15) in which the UAVs may park and/or charge their batteries.

- **Crowd Management:** Safety and security are major concerns for every smart city and they will be even more critical in future megalopolises. Already today, UAVs are playing a huge role in crowd management [8, 259, 319], and could definitely improve this field in the future. For example, police and municipal agencies can use UAVs to keep an eye on the crowd during any event (11). This will result in safer cities to live in as well but will raise privacy issues.

- **Natural Disaster Control and Emergency Response:** In the case of disasters in the megalopolis, UAVs can be used to minimize the response time and losses. Floods, fires, and earthquakes are some of the best examples in which authorities can take precautionary measures by monitoring (11) and deploying medical teams (2) or by providing communication infrastructure (12). UAVs can here analyze the entire situation and help with a quicker response than emergency calls.

As shown in this section, there is a confirmed tendency towards the development of increasingly autonomous UAV systems. This evolution would minimize the human intervention by relieving the human operator from the burden of continuously monitoring the UAVs. Nevertheless, in unpredictable situations, the UAV behavior might not conform to the expectations of the human operator. For instance, in a product delivery scenario, an autonomous UAV may choose to deviate from its expected path because of an unforeseen event. Enhancing the UAVs with explaining capabilities would allow the human operator to understand the reasons behind UAV behavior and raises its trust in the autonomous UAV system. Furthermore, developing explainable UAVs would have a very positive impact on human-machine teaming. This thesis has tackled the explainability in the domain of goal-driven systems such as UAVs. However, as this is in its early stages, some challenges or research directions may arise and the next section discusses the most related ones to the thesis.

10.2.2/ Research Directions: Explainability

The following list provides a synthesis of the paramount research directions related to the domain of XAI.

10.2.2.1/ User-aware Explainable Artificial Intelligence

Humans have different cognitive capabilities, and hence different cognitive loads to handle the explanations. Accordingly, HAExA can be extended to be not only context-aware but also user-aware. This allows for generating even more adaptive explanations and in particular, this makes adaptive filtering more personalized to the human user by calculating personalized human cognitive loads. Measuring the human cognitive load and its effects on the understandability of humans is a hot topic now with very recent research (*e.g.* [80]).

A user-aware XAI system may facilitate several benefits. Firstly, the involvement of the human could increase, and hence the direction of interactive explanations could be also considered, *i.e.* the feedback provided by the human could be integrated into the proposed architecture where the human becomes a human-in-the-loop. Verstaevel et al. [295] have presented a MAS able to dynamically learn and reuse contexts from demonstrations performed by a human tutor. This knowledge could later be used to formulate explanations. Creating a model of the user as a part of the explanation reception phase should be investigated. For this point, both the explanations generation and communication sub-processes will be affected, *i.e.* the proposed architecture will be not only context-aware but also user-aware. The direction of interactive explanations could be also explored, especially that there are almost no agent architectures among the most popular ABS frameworks enabling to easily model human actors [39]. Secondly, the trust in XAI systems may increase, as we have seen other works (*e.g.* [185]) that achieving the trust of humans in XAI systems remains a challenge and some very recent works have started to focus on enhancing human trust by explaining robot behavior in their work [83]. For that, more experiments should be conducted, *e.g.* using virtual and social agents. Thirdly, a metric or measure of explanation human cognitive load that is related to the explanation reception could be derived from such user-aware XAI systems. This can be investigated by empirical evidence or by designing a mathematical approximation akin to the law of diminishing marginal utility.

Some works have started to tackle the challenge of building a model that considers the preferences of the human, *e.g.* providing explanations depending on the user age [141], or personalized recommendations based on the user preferences [235]. Previous works considered user knowledge (*e.g.* [200]). They classified a user as a beginner or expert and used this to provide explanations that better suit the preferences of the human.

However, more elaborate user modeling is required for good personalized explanations. Abdulrahman et al. [3] also started to tackle the challenge of user-aware models in their very recent works [3, 2]. However, the proposals and the evaluations are still in their first steps, and relatively little research has been conducted regarding personalized explanations [16].

In such a context, extending HAExA with a system that facilitates the communication between humans, agents, and robots is interesting. The Human-Agent-Robot-Machine-Sensor (HARMS) model for interactions among heterogeneous actors [189] is a good candidate for such a task. HARMS connects actors such that all of them are indistinguishable in terms of which type of actor (*e.g.* robot, software agent, or even human) sends a message [190]. HARMS also includes other actors like machines and sensors that could be necessary and vital to consider when moving from the simulated environment to the real world [147, 146, 201].

10.2.2.2/ Verifying and Validating the Explainable Artificial Intelligence Systems

According to Torens et al. [286], the more complex a software algorithm gets, the more difficult it becomes to test. Furthermore, functional requirements are only one aspect of a system. Beyond the pure verification of a requirement lies the benchmark of the implemented solution. The resulting outcome may be determined by a test of the requirement, but the specific path to the solution can have different levels of quality. Therefore, additional tests have to verify that the specified explainability boundaries, as well as additional constraints, are met by this algorithm. Due to the lack of benchmarks in the evaluation of the domain of XAI, researchers are forced to rely on self-built mechanisms to validate their works (*e.g.* [288, 185, 166]). To be able to assess highly automated functions and to be able to assure high-quality software systems, it is necessary to implement a scoring system or a benchmark to evaluate the autonomy using non-functional requirements.

It is important to notice that benchmarks are problem-specific, and not implementation-specific. This enables developers not only to test an explainability algorithm automatically, without a manual review from an expert but also to evaluate algorithms and compare them with different implementations and solution approaches. The development of such automatic tests and benchmarks are gaining more interest for additional challenges, such as using natural language processing [177, 188], or ontologies to generate the explanations [213, 133, 69, 27, 310]. However, most of these efforts are focusing on data-driven XAI and not goal-driven XAI [246].

The growing pressure to innovate and the demand for shorter development cycles require changes in the development methodology. As a result, there is a shift in the demands of

testbed systems. This desire for shorter development times stands as opposed to the growing complexity required for developing increasingly automated and autonomous systems [227]. Enabling early validation of such system designs requires the simulation of components. This requires the development of an adapted simulation environment, possibly real-time or mixed reality simulator, composed of a collection of reusable modules combining real and virtual components (also called XiL: X-in-the-Loop, where X meaning alternatively Model/Software/Hardware and Human). In general, researchers think that the best experimentation includes 4 consecutive steps: Simulation, mixed reality, controlled environment, and open world.

Some works are appearing in the direction of providing benchmarking assistant tools for XAI. Hoffman et al. [132] have outlined in detail a bunch of metrics to evaluate XAI systems. Additionally, the National Institute of Standards and Technology in the USA has started a very recent endeavor [1] to establish the efforts of researchers in the domain of XAI to forge a new standard about explainability [232].

10.2.3/ Research Directions: Artificial Intelligence in the Domain of Unmanned Aerial Vehicles

The following list outlines a synthesis of the major challenges and research directions related to UAVs for AI in general and agents in particular.

10.2.3.1/ Integrating Unmanned Aerial Vehicles into Smart Cities

A smart city integrates heterogeneous connected objects to automate or simplify the autonomous and transparent accomplishment of various daily tasks, both personal and professional [106]. A smart city is an urban area that uses different types of electronic data collection sensors to supply information that is used to manage assets and resources efficiently. This includes data collected from citizens, devices, and assets that are processed and analyzed to monitor and manage traffic and transportation systems, power plants, water supply networks, waste management, law enforcement, information systems, schools, libraries, hospitals, and other community services [193]. The smart city concept integrates Information and Communication Technology (ICT) and various physical devices connected to the network to optimize the efficiency of city operations and services and connect to citizens [231]. Guastella et al. [112] have introduced a MAS to estimate missing information in smart environments. The goal of their work is to give, anytime and everywhere, accurate information where ad hoc sensors are missing. According to Farhan et al. [87], several opportunities for UAVs uses to support a smart city

[1] Public comment period: August 17, 2020 to October 15, 2020.

exist. These opportunities will be very beneficial to any smart city that would utilize UAVs for their economic growth and development.

One of the new trends in civilian UAV applications in smart cities is using UAVs in geospatial surveying. The main design of a smart city requires the optimization of data flows provided by wireless sensor networks as sensors are the main component of any autonomous system such as those involving UAVs. This combination of technologies creates a wide range of applications and opportunities such as fire management in open areas where the use of UAVs and micro-UAVs is very beneficial. The potentials vary from a wide range of available solutions and innovations that are evolving quickly. Due to the reliability of most UAV designs, integration of such technologies make it possible to install wireless sensors on-board to make the UAVs usable in geospatial, land surveying, and Geographic Information System (GIS) applications in smart cities in addition to being helpful for environmental analysis. These opportunities may lead to cost reduction and cutting down on the number of manpower hours involved in such activities.

The integration of UAV solutions with machine-to-machine, Radio-frequency identification (RFID) and live video streaming expanded the role of UAVs in public safety zones. This new trend will move the urban management personnel from being reactive to proactive. Moreover, the inclusion of UAVs in surveillance activities will reduce expenses and increase the efficiency of the tasks. The efficiency of safety and security systems in a city has become a genuine concern not only for smart cities but also for any type of urban community. The involvement of UAVs in smart policing activities has lately been supported by the USA Congress and top-level federal agencies such as the Bureau of Justice Assistance, and the USA Department of Justice.

UAVs can act as a third-party technology to coordinate information from various systems within a smart city. Since they are controlled at the ground station once they receive information the ground system can send commands to UAVs to direct the information to another system or UAVs.

In addition to integrating UAVs into smart cities, integrating them into large, open, and isolated space is another research direction. Applications in agriculture, environment preservation, and sustainable energy (*e.g.* solar farms monitoring, electric transport, *etc*.) are countless.

10.2.3.2/ Proposing and Evaluating the Regulations of Unmanned Aerial Vehicles

Regulations about introducing UAVs, including air traffic regulations, landing/taking off regulations, *etc*. are not yet fully developed, and there exist serious safety and privacy problems mostly due to the lack of verification/validation frameworks for regulations.

Globally, regulations regarding UAVs are still immature. So far, at the multilateral level, the International Civil Aviation Organization (ICAO) is the lead platform for framing regulations for UAV operations. Several regulations were passed to regulate the use of UAVs; however, no proposals were made from a technological point of view. Moreover, legislation varies from region to region and between countries [57].

Recognizing the enormous potential growth of UAVs, the European Aviation Safety Agency (EASA) has been tasked by the European Commission to frame regulations for UAV operations. In 2014, the Commission published "A new era for aviation—Opening the aviation market to the civilian use of remotely piloted aircraft systems safely and sustainably" [291] but this does not include the UAVs. The EASA published a comprehensive proposal in May 2017 covering the technical and operational aspects of operating UAV. According to the proposal, all UAVs above 250 g need to be registered. The proposal put the alignment of different national UAV legislations as one of Horizon 2020 goals. However, these goals have been postponed until 2050. Moreover, different European countries have different regulations – for instance, one can fly UAVs commercially in Switzerland if line-of-sight can be ensured, within certain altitude limitations, and not flying near protected areas such as airports. On the other hand, France has more restrictive regulations in place where it is mandatory that any UAV operation over a city needs to be authorized by aviation authorities. In Belgium, Brussels is planning to create "Uspace" in 2019, a European controlled space for UAVs flying above 150 m in height and weighing less than 150 kg.

In India, some work examines civilian UAV operations and analyses the major policy gaps in the country's evolving policy framework. It argues that ad hoc measures taken by agencies have been ineffective, whether in addressing issues of quality control, or response mechanisms, questions of privacy and trespass, air traffic, and legal liability [237].

USA has by far the most mature civilian UAV regulations in place. The New Small UAS Rule (107) of the Federal Aviation Administration (FAA) that was issued in August 2016 regulates most of the UAV operations, especially those related to commercial or civilian purposes. The FAA has relaxed the regulations for UAV operations in the commercial sector considering that the UAV applications are estimated to generate an additional US$82 billion for the economy of the USA.

Making progress on the issue of reaching common legislation will be a complicated task. This is because international conventions on international civilian aviation, such as the Chicago Convention, apply only to civilian manned aircraft but not to unmanned ones [57]. It is necessary to have legislation that will be open and generic in the technical aspects. This is because legislations that are limited to specific aircraft types or only permit the use of remote controls with certain characteristics would become obsolete soon, as new advances in the field of UAVs appear. Furthermore, regulations should not only consider

the civilian liability of these devices but also aspects that will assure the security of the citizens, for example, the protection of data in deployed vehicles [57]. We anticipate that the simulation, in general, and the ABS, in particular, of AI systems will help in directing the efforts to propose, evaluate, and forge new UAVs regulations.

BIBLIOGRAPHY

[1] ABAR, Sameera ; THEODOROPOULOS, Georgios K. ; LEMARINIER, Pierre ; O'HARE, Gregory M.: **"Agent Based Modelling and Simulation tools: A review of the state-of-art software"**. In *Computer Science Review* (2017)

[2] ABDULRAHMAN, Amal ; RICHARDS, Deborah: **"Modelling Therapeutic Alliance using a User-aware Explainable Embodied Conversational Agent to Promote Treatment Adherence"**. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pages 248–251

[3] ABDULRAHMAN, Amal ; RICHARDS, Deborah ; RANJBARTABAR, Hedieh ; MASCARENHAS, Samuel: **"Belief-based Agent Explanations to Encourage Behaviour Change"**. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pages 176–178

[4] ADADI, Amina ; BERRADA, Mohammed: **"Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)"**. In *IEEE Access* 6 (2018), pages 52138–52160

[5] ADAM, Carole ; GAUDOU, Benoit: **"BDI agents in social simulations: a survey"**. In *The Knowledge Engineering Review* 31 (2016), number 3, pages 207–238

[6] ADAMS, Barbara D. ; BRUYN, Lora E. ; HOUDE, Sébastien: *Trust in Automated Systems, Literature Review*. Humansystems Incorporated, 2003

[7] AGOGINO, Adrian K. ; HOLMESPARKER, Chris ; TUMER, Kagan: **"Evolving large scale UAV communication system"**. In *Genetic and Evolutionary Computation Conference, GECCO '12, Philadelphia, PA, USA, July 7-11, 2012*, 2012, pages 1023–1030. DOI: 10.1145/2330163.2330306

[8] ALABDULKARIM, Lamia ; ALRAJHI, Wafa ; ALOBOUD, Ebtesam: **"Urban Analytics in Crowd Management in the Context of Hajj"**. In *Int. Conf. on Social Computing and Social Media* Springer (event), 2016, pages 249–257

[9] ALBANI, Dario ; MANONI, Tiziano ; NARDI, Daniele ; TRIANNI, Vito: **"Dynamic UAV Swarm Deployment for Non-Uniform Coverage"**. In *Proc. of 17th Int. Conf. on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, 2018, pages 523–531

[10] ALBAUM, Gerald: **"The Likert scale revisited"**. In *Market Research Society. Journal.* 39 (1997), number 2, pages 1–21

[11] ALLAN, Robert J.: *Survey of agent based modelling and simulation tools*. Science & Technology Facilities Council, 2010

[12] ALONSO-MORA, Javier ; SAMARANAYAKE, Samitha ; WALLAR, Alex ; FRAZZOLI, Emilio ; RUS, Daniela: **"On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment"**. In *National Academy of Sciences* 114 (2017), number 3, pages 462–467

[13] AMUKELE, Timothy ; NESS, Paul M. ; TOBIAN, Aaron A. ; BOYD, Joan ; STREET, Jeff: **"Drone transportation of blood products"**. In *Transfusion* 57 (2017), number 3, pages 582–588

[14] ANDERSON, John R. ; MATESSA, Michael ; LEBIERE, Christian: **"ACT-R: A theory of higher level cognition and its relation to visual attention"**. In *Human–Computer Interaction* 12 (1997), number 4, pages 439–462

[15] ANJOMSHOAE, Sule ; FRÄMLING, Kary ; NAJJAR, Amro: **"Explanations of black-box model predictions by contextual importance and utility"**. In *International Workshop on eXplainable, Transparent Autonomous Agents and Multi-Agent Systems* Springer (event), 2019, pages 95–109

[16] ANJOMSHOAE, Sule ; NAJJAR, Amro ; CALVARESI, Davide ; FRÄMLING, Kary: **"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 (event), 2019, pages 1078–1088

[17] AROKIASAMI, Willson A. ; VADAKKEPAT, Prahlad ; TAN, Kay C. ; SRINIVASAN, Dipti: **"Interoperable multi-agent framework for unmanned aerial/ground vehicles: towards robot autonomy"**. In *Complex & Intelligent Systems* 2 (2016), number 1, pages 45–59

[18] ASHRAF, Adnan ; MAJD, Amin ; TROUBITSYNA, Elena: **"Towards a realtime, collision-free motion coordination and navigation system for a UAV fleet"**. In *Proc. of Fifth European Conf. on the Engineering of Computer-Based Systems, ECBS 2017, Larnaca, Cyprus, August 31 - September 01*, 2017, pages 11:1–11:9. DOI: 10.1145/3123779.3123805

[19] AZARIA, Amos ; FIOSINA, Jelena ; GREVE, Maike ; HAZON, Noam ; KOLBE, Lutz ; LEMBCKE, Tim-Benjamin ; MÜLLER, Jörg P ; SCHLEIBAUM, Sören ; VOLLRATH, Mark: **"AI for Explaining Decisions in Multi-Agent Environments"**. In *arXiv preprint arXiv:1910.04404* (2019)

[20] AZOULAY, Rina ; RECHES, Shulamit: **"UAV Flocks Forming for Crowded Flight Environments"**. In *Proc. of 11th Int. Conf. on Agents and Artificial Intelligence, ICAART 2019, Volume 2*, URL https://doi.org/10.5220/0007369401540163, 2019, pages 154–163. DOI: 10.5220/0007369401540163

[21] BADEIG, Fabien ; BALBO, Flavien ; PINSON, Suzanne: **"A Contextual Environment Approach for Multi-agent-based Simulation."**. In *ICAART (2)*, 2010, pages 212–217

[22] BADEIG, Fabien ; BALBO, Flavien ; ZARGAYOUNA, Mahdi: **"Dynamically Configurable Multi-agent Simulation for Crisis Management"**. In *Agents and Multi-agent Systems: Technologies and Applications 2019*. Springer, 2020, pages 343–352

[23] BAINBRIDGE, Wilma A. ; HART, Justin ; KIM, Elizabeth S. ; SCASSELLATI, Brian: **"The effect of presence on human-robot interaction"**. In *RO-MAN 17th IEEE Int. Symposium on Robot and Human Interactive Communication*, 2008, pages 701–706

[24] BAMBURRY, Dane: **"Drones: Designed for product delivery"**. In *Design Management Review* 26 (2015), number 1, pages 40–48

[25] BARAKA, Kim ; PAIVA, Ana ; VELOSO, Manuela: **"Expressive lights for revealing mobile service robot state"**. In *Robot 2015: Second Iberian Robotics Conference* Springer (event), 2016, pages 107–119

[26] BAYAT, Behzad ; CRASTA, Naveena ; CRESPI, Alessandro ; PASCOAL, António M ; IJSPEERT, Auke: **"Environmental monitoring using autonomous vehicles: a survey of recent searching techniques"**. In *Current opinion in biotechnology* 45 (2017), pages 76–84

[27] BÉHÉ, Florian ; GALLAND, Stéphane ; GAUD, Nicolas ; NICOLLE, Christophe ; KOUKAM, Abderrafiaa: **"An ontology-based metamodel for multiagent-based simulations"**. In *Simulation Modelling Practice and Theory* 40 (2014), pages 64–85

[28] BEHKAMAL, Behshid ; KAHANI, Mohsen ; AKBARI, Mohammad K.: **"Customizing ISO 9126 quality model for evaluation of B2B applications"**. In *Information and software technology* 51 (2009), number 3, pages 599–609

[29] BEKEY, George A.: *Autonomous robots: from biological inspiration to implementation and control*. MIT press, 2005

[30] BELLIFEMINE, Fabio L. ; CAIRE, Giovanni ; GREENWOOD, Dominic: *Developing multi-agent systems with JADE*. Volume 7. John Wiley & Sons, 2007

[31] BENEDETTI, Massimiliano D. ; D'URSO, Fabio ; MESSINA, Fabrizio ; PAPPALARDO, Giuseppe ; SANTORO, Corrado: **"UAV-based Aerial Monitoring: A Performance Evaluation of a Self-Organising Flocking Algorithm"**. In *10th Int. Conf. on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC 2015, Krakow, Poland, November 4-6*, 2015, pages 248–255. DOI: 10.1109/3PGCIC.2015.78

[32] BENTZ, William ; PANAGOU, Dimitra: **"3D dynamic coverage and avoidance control in power-constrained UAV surveillance networks"**. In *Unmanned Aircraft Systems (ICUAS), 2017 Int. Conf. on* IEEE (event), 2017, pages 1–10

[33] BETHEL, Cindy L.: *Robots Without Faces: Non-verbal Social Human-robot Interaction*. Tampa, FL, USA, PhD Thesis, 2009. – AAI3420462

[34] BIRAN, Or ; COTTON, Courtenay: **"Explanation and justification in machine learning: A survey"**. In *IJCAI-17 workshop on explainable AI (XAI)* Volume 8, 2017

[35] BLAIKIE, Norman: *Analyzing quantitative data: From description to explanation*. Sage, 2003

[36] BLUMER, Anselm ; EHRENFEUCHT, Andrzej ; HAUSSLER, David ; WARMUTH, Manfred K.: **"Occam's razor"**. In *Information processing letters* 24 (1987), number 6, pages 377–380

[37] BORDINI, Rafael H. ; HÜBNER, Jomi F. ; WOOLDRIDGE, Michael: *Programming multi-agent systems in AgentSpeak using Jason*. Volume 8. John Wiley & Sons, 2007

[38] BORGO, Rita ; CASHMORE, Michael ; MAGAZZENI, Daniele: **"Towards providing explanations for AI planner decisions"**. In *arXiv preprint arXiv:1810.06338* (2018)

[39] BOURGAIS, Mathieu ; TAILLANDIER, Patrick ; VERCOUTER, Laurent: **"BEN: An Agent Architecture for Explainable and Expressive Behavior in Social Simulation"**. In *International Workshop on eXplainable, TRansparent Autonomous Agents and Multi-Agent Systems* Springer (event), 2019, pages 147–163

[40] BRADLEY, Graham L. ; SPARKS, Beverley A.: **"Dealing with service failures: The use of explanations"**. In *Journal of Travel & Tourism Marketing* 26 (2009), number 2, pages 129–143

[41] BRATMAN, Michael ; OTHERS: *Intention, plans, and practical reason*. Volume 10. Harvard University Press Cambridge, MA, 1987

[42] BRATMAN, Michael E. ; ISRAEL, David J. ; POLLACK, Martha E.: **"Plans and resource-bounded practical reasoning"**. In *Computational intelligence* 4 (1988), number 3, pages 349–355

[43] BRERETON, Pearl ; KITCHENHAM, Barbara A. ; BUDGEN, David ; TURNER, Mark ; KHALIL, Mohamed: **"Lessons from applying the systematic literature review process within the software engineering domain"**. In *Journal of systems and software* 80 (2007), number 4, pages 571–583

[44] BROEKENS, Joost ; DEGROOT, Doug ; KOSTERS, Walter A.: **"Formal models of appraisal: Theory, specification, and computational model"**. In *Cognitive Systems Research* 9 (2008), number 3, pages 173–197

[45] BROEKENS, Joost ; HARBERS, Maaike ; HINDRIKS, Koen ; VAN DEN BOSCH, Karel ; JONKER, Catholijn ; MEYER, John-Jules: **"Do you get it? User-evaluated explainable BDI agents"**. In *German Conf. on Multiagent System Technologies* Springer (event), 2010, pages 28–39

[46] BUDGEN, David ; BRERETON, Pearl: **"Performing systematic literature reviews in software engineering"**. In *Proc. of 28th Int. Conf. on Software engineering* ACM (event), 2006, pages 1051–1052

[47] BÜRKLE, Axel: **"Collaborating miniature drones for surveillance and reconnaissance"**. In *Proc. of SPIE Vol* Volume 7480, 2009, pages 74800H–1

[48] BÜRKLE, Axel ; LEUCHTER, Sandro: **"Development of Micro UAV Swarms"**. In *Autonome Mobile Systeme 2009 - 21. Fachgespräch, Karlsruhe*, Springer, 2009, pages 217–224. DOI: 10.1007/978-3-642-10284-4_28

[49] BÜRKLE, Axel ; SEGOR, Florian ; KOLLMANN, Matthias: **"Towards Autonomous Micro UAV Swarms"**. In *Journal of Intelligent and Robotic Systems* 61 (2011), number 1-4, pages 339–353. DOI: 10.1007/s10846-010-9492-x

[50] CALVARESI, Davide ; CESARINI, Daniel ; SERNANI, Paolo ; MARINONI, Mauro ; DRAGONI, Aldo F. ; STURM, Arnon: **"Exploring the ambient assisted living domain: a systematic review"**. In *Journal of Ambient Intelligence and Humanized Computing* 8 (2017), number 2, pages 239–257

[51] CALVARESI, Davide ; MUALLA, Yazan ; NAJJAR, Amro ; GALLAND, Stéphane ; SCHUMACHER, Michael: **"Explainable Multi-Agent Systems through Blockchain Technology"**. In *Proc. of 1st Int. Workshop on eXplainable TRansparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS 2019)*, 2019

[52] CAREY, Peter: *Data protection: a practical guide to UK and EU law*. Oxford University Press, Inc., 2018

[53] CASTLE, Christian J. ; CROOKS, Andrew T.: *Principles and Concepts of Agent-Based Modelling for Developing Geospatial Simulations*. Centre for Advanced Spatial Analysis, University College London (UCL). 2006

[54] CELLIER, Francois E. ; GREIFENEDER, Jurgen: *Continuous System Modeling*. Springer, 1991 DOI: 10.1007/978-1-4757-3922-0

[55] CHAKRABORTI, Tathagata ; FADNIS, Kshitij P. ; TALAMADUPULA, Kartik ; DHOLAKIA, Mishal ; SRIVASTAVA, Biplav ; KEPHART, Jeffrey O. ; BELLAMY, Rachel K.: **"Visualizations for an Explainable Planning Agent."**. In *IJCAI*, 2018, pages 5820–5822

[56] CHAKRABORTI, Tathagata ; SREEDHARAN, Sarath ; ZHANG, Yu ; KAMBHAMPATI, Subbarao: **"Plan explanations as model reconciliation: Moving beyond explanation as soliloquy"**. In *arXiv preprint arXiv:1701.08317* (2017)

[57] CHAMOSO, Pablo ; GONZÁLEZ-BRIONES, Alfonso ; RIVAS, Alberto ; BUENO DE MATA, Federico ; CORCHADO, Juan M.: **"The Use of Drones in Spain: Towards a Platform for Controlling UAVs in Urban Environments"**. In *Sensors* 18 (2018), number 5, pages 1416

[58] CHANDRASEKARAN, Arjun ; YADAV, Deshraj ; CHATTOPADHYAY, Prithvijit ; PRABHU, Viraj ; PARIKH, Devi: **"It Takes Two to Tango: Towards Theory of AI's Mind"**. In *arXiv preprint arXiv:1704.00717* (2017)

[59] CHEN, Yueyue ; ZHANG, Haidong ; XU, Ming: **"The coverage problem in UAV network: A survey"**. In *Computing, Communication and Networking Technologies (ICCCNT), 2014 Int. Conf. on* IEEE (event), 2014, pages 1–5

[60] CHIN-PARKER, Seth ; CANTELON, Julie: **"Contrastive Constraints Guide Explanation-Based Category Learning"**. In *Cognitive science* 41 (2017), number 6, pages 1645–1655

[61] CHMAJ, Grzegorz ; SELVARAJ, Henry: **"Distributed processing applications for UAV/drones: a survey"**. In *Progress in Systems Engineering*. Springer, 2015, pages 449–454

[62] CHOMSKY, Noam ; COLLINS, Chris: *Beyond explanatory adequacy*. Volume 20. mitwpl, 2001

[63] CHURCHLAND, Paul M.: **"Folk psychology and the explanation of human behavior"**. In *Philosophical Perspectives* 3 (1989), pages 225–241

[64] CIARLETTA, Laurent ; GUENARD, Adrien ; PRESSE, Yannick ; GALTIER, Virgine ; SONG, Ye-Qiong ; PONSART, Jean-Christophe ; ABERKANE, Samir ; THEILLIOL, Didier: **"Simulation and Platform Tools to develop safe flock of UAVs: a CPS**

Application-Driven Research". In *Unmanned Aircraft Systems (ICUAS), 2014 Int. Conf. on* IEEE (event), 2014, pages 95–102

[65] CIMINO, Mario G. C. A. ; LAZZERI, Alessandro ; VAGLINI, Gigliola: **"Combining stigmergic and flocking behaviors to coordinate swarms of drones performing target search"**. In *6th Int. Conf. on Information, Intelligence, Systems and Applications, IISA 2015, Corfu, Greece, July 6-8*, 2015, pages 1–6. DOI: 10.1109/IISA.2015.7387990

[66] CLOUGH, Bruce T.: **"Metrics, schmetrics! How the heck do you determine a UAV's autonomy anyway"** / AIR FORCE RESEARCH LAB WRIGHT-PATTERSON AFB OH. 2002. – Research Report

[67] COHEN, Louis ; MANION, Lawrence ; MORRISON, Keith: *Research methods in education*. routledge, 2002

[68] COLLIER, Nick: **"Repast: An extensible framework for agent simulation"**. In *The University of Chicago's Social Science Research* 36 (2003)

[69] CONFALONIERI, Roberto ; PRADO, Fermín M. del ; AGRAMUNT, Sebastia ; MALAGARRIGA, Daniel ; FAGGION, Daniele ; WEYDE, Tillman ; BESOLD, Tarek R.: **"An Ontology-based Approach to Explaining Artificial Neural Networks"**. In *arXiv preprint arXiv:1906.08362* (2019)

[70] CONTRERAS, Heles: **"Simplicity, descriptive adequacy, and binary features"**. In *Language* (1969), pages 1–8

[71] CRAIGHEAD, Jeff ; MURPHY, Robin ; BURKE, Jenny ; GOLDIEZ, Brian: **"A survey of commercial & open source unmanned vehicle simulators"**. In *Robotics and Automation, 2007 IEEE Int. Conf. on* IEEE (event), 2007, pages 852–857

[72] D'ANDREA, Raffaello: **"Guest editorial can drones deliver?"**. In *IEEE Transactions on Automation Science and Engineering* 11 (2014), number 3, pages 647–648

[73] DE BENEDETTI, Massimiliano ; D'URSO, Fabio ; MESSINA, Fabrizio ; PAPPALARDO, Giuseppe ; SANTORO, Corrado: **"Self-Organising UAVs for Wide Area Fault-tolerant Aerial Monitoring"**. In *Proc. of 16th Workshop "From Objects to Agents", Naples, Italy, June 17-19*, 2015, pages 135–141

[74] DE BENEDETTI, Massimiliano ; D'URSO, Fabio ; MESSINA, Fabrizio ; PAPPALARDO, Giuseppe ; SANTORO, Corrado: **"3D Simulation of Unmanned Aerial Vehicles"**. In *Proc. of 18th Workshop "From Objects to Agents", Scilla (RC), Italy, June 15-16*, 2017, pages 7–12

[75] DENNETT, Daniel C.: **"Three kinds of intentional psychology"**. In *Perspectives in the philosophy of language: A concise anthology* (1978), pages 163–186

[76] DENNIS, Louise A. ; FARWER, Berndt: **"Gwendolen: a BDI language for verifiable agents"**. In *Proc. of AISB 2008 Symposium on Logic and the Simulation of Interaction and Reasoning, Society for the Study of Artificial Intelligence and Simulation of Behaviour*, 2008, pages 16–23

[77] DHURANDHAR, Amit ; IYENGAR, Vijay ; LUSS, Ronny ; SHANMUGAM, Karthikeyan: **"TIP: Typifying the Interpretability of Procedures"**. In *CoRR* abs/1706.02952 (2017). – URL http://arxiv.org/abs/1706.02952

[78] DIAS, Paulo S. ; FRAGA, Sergio L. ; GOMES, Rui M. ; GONCALVES, Gil M. ; PEREIRA, Fernando L. ; PINTO, Jose ; SOUSA, Joao B.: **"Neptus-a framework to support multiple vehicle operation"**. In *Oceans 2005-Europe* Volume 2 IEEE (event), 2005, pages 963–968

[79] DORAN, Derek ; SCHULZ, Sarah ; BESOLD, Tarek R.: **"What does explainable AI really mean? A new conceptualization of perspectives"**. In *arXiv preprint arXiv:1710.00794* (2017)

[80] DOUGHERTY, Sean: **"Partnering People with Deep Learning Systems: Human Cognitive Effects of Explanations"**. (2019)

[81] DROMEY, R. G.: **"A model for software product quality"**. In *IEEE Transactions on Software Engineering* 21 (1995), number 2, pages 146–162

[82] DZINDOLET, Mary T. ; PETERSON, Scott A. ; POMRANKY, Regina A. ; PIERCE, Linda G. ; BECK, Hall P.: **"The role of trust in automation reliance"**. In *Int. journal of human-computer studies* 58 (2003), number 6, pages 697–718

[83] EDMONDS, Mark ; GAO, Feng ; LIU, Hangxin ; XIE, Xu ; QI, Siyuan ; ROTHROCK, Brandon ; ZHU, Yixin ; WU, Ying N. ; LU, Hongjing ; ZHU, Song-Chun: **"A tale of two explanations: Enhancing human trust by explaining robot behavior"**. In *Science Robotics* 4 (2019), number 37

[84] EPSTEIN, Susan L.: **"For the right reasons: The FORR architecture for learning in a skill domain"**. In *Cognitive science* 18 (1994), number 3, pages 479–511

[85] EVERTSZ, Rick ; THANGARAJAH, John ; LY, Thanh: **"A BDI-Based Methodology for Eliciting Tactical Decision-Making Expertise"**. In SARKER, Ruhul (editors) ; ABBASS, Hussein A. (editors) ; DUNSTALL, Simon (editors) ; KILBY, Philip (editors) ; DAVIS, Richard (editors) ; YOUNG, Leon (editors): *Data and Decision Sciences in Action*. Cham : Springer International Publishing, 2018, pages 13–26. – ISBN 978-3-319-55914-8

[86] EVERTSZ, Rick ; THANGARAJAH, John ; YADAV, Nitin ; LY, Thanh: **"A framework for modelling tactical decision-making in autonomous systems"**. In *Journal of Systems and Software* 110 (2015), pages 222–238. DOI: 10.1016/j.jss.2015.08.046

[87] FARHAN, Mohammed ; IDRIES, Ahmed ; NADER, Mohamed ; AL-JAROODI, Jameela ; JAWHAR, Imad: **"UAVs for Smart Cities: Opportunities and Challenges"**. In *Int. Conf. on Unmanned Aircraft Systems (ICUAS)*. Orlando, FL, USA, May 2014

[88] FAWAZ, Wissam ; ABOU-RJEILY, Chadi ; ASSI, Chadi: **"UAV-aided cooperation for FSO communication systems"**. In *IEEE Communications Magazine* 56 (2018), number 1, pages 70–75

[89] FERBER, Jacques: *Multi-agent systems: an introduction to distributed artificial intelligence*. Volume 1. Addison-Wesley Reading, 1999

[90] FERRAG, Abdelmoumen ; OUSSAR, Abdelatif ; GUIATNI, Mohamed: **"Robust coordinated motion planning for UGV/UAV agents in disturbed environment"**. In *Modelling, Identification and Control (ICMIC), 2016 8th Int. Conf. on* IEEE (event), 2016, pages 472–477

[91] FININ, Tim ; FRITZSON, Richard ; MCKAY, Don ; MCENTIRE, Robin: **"KQML as an agent communication language"**. In *Proceedings of the third international conference on Information and knowledge management*, 1994, pages 456–463

[92] FISCHER, Joel E. ; GREENHALGH, Chris ; JIANG, Wenchao ; RAMCHURN, Sarvapali D. ; WU, Feng ; RODDEN, Tom: **"In-the-loop or on-the-loop? Interactional arrangements to support team coordination with a planning agent"**. In *Concurrency and Computation: Practice and Experience* (2017), pages e4082

[93] FITZPATRICK, Ronan: **"Software quality: definitions and strategic issues"**. In *Reports* (1996), pages 1

[94] FLOREANO, Dario ; WOOD, Robert J.: **"Science, technology and the future of small autonomous drones"**. In *Nature* 521 (2015), number 7553, pages 460–466

[95] FLOYD, Michael W. ; AHA, David W.: **"Incorporating transparency during trust-guided behavior adaptation"**. In *International Conference on Case-Based Reasoning* Springer (event), 2016, pages 124–138

[96] FOX, Maria ; LONG, Derek ; MAGAZZENI, Daniele: **"Explainable planning"**. In *arXiv preprint arXiv:1709.10256* (2017)

[97] FRANKLIN, Stan ; PATTERSON JR, FG: **"The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent"**. In *pat* 703 (2006), pages 764–1004

[98] FULFORD, CD ; LIE, NHM ; EARON, EJP ; HUQ, R ; RABBATH, CA: **"The Vehicle Abstraction Layer: A simplified approach to multi-agent, autonomous UAV systems development"**. In *System Simulation and Scientific Computing, 2008. ICSC 2008. Asia Simulation Conference-7th Int. Conf. on* IEEE (event), 2008, pages 483–487

[99] GALLAND, Stéphane ; GAUD, Nicolas: **"Organizational and Holonic Modelling of a Simulated and Synthetic Spatial Environment"**. In *In E4MAS 2014 - 10 years later, LNCS* Volume 9068, 2015, pages 1–23. DOI: 10.1007/978-3-319-23850-0

[100] GALLAND, Stéphane ; GAUD, Nicolas ; DEMANGE, Jonathan ; KOUKAM, Abderrafiaa: **"Environment model for multiagent-based simulation of 3D urban systems"**. In *the 7th European Workshop on Multiagent Systems (EUMAS09), Ayia Napa, Cyprus*, 2009

[101] GALLAND, Stéphane ; RODRIGUEZ, Sebastian ; GAUD, Nicolas: **"Run-time Environment for the SARL Agent-Programming Language: the Example of the Janus platform"**. In *Int. Journal on Future Generation Computer Systems* (2017), October. – ISSN 0167-739X. DOI: 10.1016/j.future.2017.10.020

[102] GALSTER, Matthias ; WEYNS, Danny ; TOFAN, Dan ; MICHALIK, Bartosz ; AVGERIOU, Paris: **"Variability in software systems—a systematic literature review"**. In *IEEE Transactions on Software Engineering* 40 (2014), number 3, pages 282–306

[103] GEORGÉ, Jean-Pierre ; GLEIZES, Marie-Pierre ; GLIZE, Pierre: **"Conception de systèmes adaptatifs à fonctionnalité émergente: la théorie Amas."**. In *Revue d'intelligence artificielle* 17 (2003), number 4, pages 591–626

[104] GHALLAB, Malik ; NAU, Dana ; TRAVERSO, Paolo: *Automated Planning: theory and practice*. Elsevier, 2004

[105] GILBERT, Nigel ; CONTE, Rosaria ; OTHERS: *Artificial societies: The computer simulation of social life*. Routledge, 2006

[106] GLASER, Alex ; ALLMENDINGER, Glen: **"Smart Cities Growth Opportunities Overview 2016–2021"**. Boulder, USA, 2017. – Research Report

[107] GOEDERTIER, Stijn ; MUES, Christophe ; VANTHIENEN, Jan: **"Specifying process-aware access control rules in SBVR"**. In *International Workshop on Rules and*

Rule Markup Languages for the Semantic Web Springer (event), 2007, pages 39–52

[108] GOEDERTIER, Stijn ; VANTHIENEN, Jan: **"Declarative process modeling with business vocabulary and business rules"**. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* Springer (event), 2007, pages 603–612

[109] GONG, Ze ; ZHANG, Yu: **"Behavior Explanation as Intention Signaling in Human-Robot Teaming"**. In *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* IEEE (event), 2018, pages 1005–1011

[110] GOODMAN, Nelson: *Problems and projects*. Bobbs-Merrill, 1972

[111] GREGORI, Miguel E. ; CÁMARA, Javier P. ; BADA, Gustavo A.: **"A jabber-based multi-agent system platform"**. In *Proc. of fifth Int. joint Conf. on Autonomous agents and multiagent systems* ACM (event), 2006, pages 1282–1284

[112] GUASTELLA, Davide A. ; CAMPS, Valérie ; GLEIZES, Marie-Pierre: **"Multi-agent Systems for Estimating Missing Information in Smart Cities."**. In *ICAART (2)*, 2019, pages 214–223

[113] GUIDOTTI, Riccardo ; MONREALE, Anna ; RUGGIERI, Salvatore ; TURINI, Franco ; GIANNOTTI, Fosca ; PEDRESCHI, Dino: **"A survey of methods for explaining black box models"**. In *ACM Computing Surveys (CSUR)* 51 (2018), number 5, pages 93

[114] GUNETTI, Paolo ; DODD, Tony J. ; THOMPSON, Haydn: **"Simulation of a Soar-Based Autonomous Mission Management System for Unmanned Aircraft"**. In *J. Aerospace Inf. Sys.* 10 (2013), number 2, pages 53–70. DOI: 10.2514/1.53282

[115] GUNETTI, Paolo ; THOMPSON, Haydn ; DODD, Tony J.: **"Autonomous mission management for UAVs using soar intelligent agents"**. In *Int. J. Systems Science* 44 (2013), number 5, pages 831–852. DOI: 10.1080/00207721.2011.626902

[116] GUNNING, David: **"Explainable artificial intelligence (XAI)"**. In *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017)

[117] GUPTA, Lav ; JAIN, Raj ; VASZKUN, Gabor: **"Survey of important issues in UAV communication networks"**. In *IEEE Communications Surveys & Tutorials* 18 (2016), number 2, pages 1123–1152

[118] GUPTA, Ritu ; KANSAL, Gaurav: **"A survey on comparative study of mobile agent platforms"**. In *Int. Journal of Engineering Science and Technology* 3 (2011), number 3

[119] HARBERS, Maaike ; BOSCH, Karel van den ; MEYER, John-Jules: **"Design and evaluation of explainable BDI agents"**. In *2010 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology* Volume 2, 2010, pages 125–132

[120] HARBERS, Maaike ; BOSCH, Karel van den ; MEYER, John-Jules C.: **"A study into preferred explanations of virtual agent behavior"**. In *Int. Workshop on Intelligent Virtual Agents* Springer (event), 2009, pages 132–145

[121] HARBERS, Maaike ; BOSCH, Karel van den ; MEYER, John-Jules C.: **"A Theoretical Framework for Explaining Agent Behavior."**. In *SIMULTECH*, SciTePress, 2011, pages 228–231

[122] HARBERS, Maaike ; BROEKENS, Joost ; VAN DEN BOSCH, Karel ; MEYER, John-Jules: **"Guidelines for developing explainable cognitive models"**. In *Proc. of ICCM* Citeseer (event), 2010, pages 85–90

[123] HASTIE, Helen ; CHIYAH GARCIA, Francisco J. ; ROBB, David A. ; LASKOV, Atanas ; PATRON, Pedro: **"MIRIAM: A Multimodal Interface for Explaining the Reasoning Behind Actions of Remote Autonomous Systems"**. In *Proceedings of the 2018 on International Conference on Multimodal Interaction* ACM (event), 2018, pages 557–558

[124] HASTIE, Helen ; LIU, Xingkun ; PATRON, Pedro: **"Trust triggers for multimodal command and control interfaces"**. In *Proc. of 19th ACM Int. Conf. on Multimodal Interaction* ACM (event), 2017, pages 261–268

[125] HAYAT, Samira ; YANMAZ, Evsen ; MUZAFFAR, Raheeb: **"Survey on Unmanned Aerial Vehicle Networks for Civil Applications: A Communications Viewpoint"**. In *IEEE Communications Surveys and Tutorials* 18 (2016), number 4, pages 2624–2661

[126] HAYNES, Steven R. ; COHEN, Mark A. ; RITTER, Frank E.: **"Designs for explaining intelligent agents"**. In *International Journal of Human-Computer Studies* 67 (2009), number 1, pages 90–110

[127] HELLSTRÖM, Thomas ; BENSCH, Suna: **"Understandable robots-what, why, and how"**. In *Paladyn, Journal of Behavioral Robotics* 9 (2018), number 1, pages 110–123

[128] HESSLOW, Germund: **"The problem of causal selection"**. In *Contemporary science and natural explanation: Commonsense conceptions of causality* (1988), pages 11–32

[129] HILTON, Denis J. ; SLUGOSKI, Ben R.: **"Knowledge-based causal attribution: The abnormal conditions focus model."**. In *Psychological review* 93 (1986), number 1, pages 75

[130] HLEG, AI: **"Ethics guidelines for trustworthy AI"**. In *B-1049 Brussels* (2019)

[131] HOFFER, Nathan V. ; COOPMANS, Calvin ; JENSEN, Austin M. ; CHEN, YangQuan: **"Small low-cost unmanned aerial vehicle system identification: a survey and categorization"**. In *Unmanned Aircraft Systems (ICUAS), 2013 Int. Conf. on* IEEE (event), 2013, pages 897–904

[132] HOFFMAN, Robert R. ; MUELLER, Shane T. ; KLEIN, Gary ; LITMAN, Jordan: **"Metrics for explainable AI: Challenges and prospects"**. In *arXiv preprint arXiv:1812.04608* (2018)

[133] HOLZINGER, Andreas ; KIESEBERG, Peter ; WEIPPL, Edgar ; TJOA, A M.: **"Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI"**. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* Springer (event), 2018, pages 1–8

[134] HOLZMANN, Gerard J.: *The SPIN model checker: Primer and reference manual*. Volume 1003. Addison-Wesley Reading, 2004

[135] INT. ORGANIZATION FOR STANDARDIZATION AND INT. ELECTROTECHNICAL COMMISSION: *Software Engineering Product Quality: Quality model*. Volume 1. ISO/IEC, 2001

[136] JAMIESON, Susan ; OTHERS: **"Likert scales: how to (ab)use them"**. In *Medical education* 38 (2004), number 12, pages 1217–1218

[137] JENNINGS, Nicholas R. ; SYCARA, Katia ; WOOLDRIDGE, Michael: **"A roadmap of agent research and development"**. In *Autonomous agents and multi-agent systems* 1 (1998), number 1, pages 7–38

[138] KAMBAYASHI, Yasushi ; YAJIMA, Hideaki ; SHYOJI, Tadashi ; OIKAWA, Ryotaro ; TAKIMOTO, Munehiro: **"Formation Control of Swarm Robots Using Mobile Agents"**. In *Vietnam J. Computer Science* 6 (2019), number 2, pages 193–222. DOI: 10.1142/S2196888819500131

[139] KAMPIK, Timotheus ; NIEVES, Juan C.: **"JS-son–A Lean, Extensible JavaScript Agent Programming Library"**. In *Engineering Multi-Agent Systems*. Cham : Springer International Publishing, 2020

[140] KANDIL, Amr A. ; WAGNER, Achim ; GOTTA, Alexander ; BADREDDIN, Essamedin: **"Collision avoidance in a recursive nested behaviour control structure for Unmanned Aerial Vehicles"**. In *Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics, Istanbul, Turkey, 10-13 October*, 2010, pages 4276–4281. DOI: 10.1109/ICSMC.2010.5642396

[141] KAPTEIN, Frank ; BROEKENS, Joost ; HINDRIKS, Koen ; NEERINCX, Mark: **"Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults"**. In *Robot and Human Interactive Communication (RO-MAN), 26th IEEE International Symposium on* IEEE (event), IEEE, 2017, pages 676–682

[142] KAPTEIN, Frank ; BROEKENS, Joost ; HINDRIKS, Koen ; NEERINCX, Mark: **"The role of emotion in self-explanations by cognitive agents"**. In *Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2017 Seventh International Conference on* IEEE (event), 2017, pages 88–93

[143] KHALEGHI, Amirreza M. ; XU, Dong ; WANG, Zhenrui ; LI, Mingyang ; LOBOS, Alfonso ; LIU, Jian ; SON, Young-Jun: **"A DDDAMS-based planning and control framework for surveillance and crowd control via UAVs and UGVs"**. In *Expert Syst. Appl.* 40 (2013), number 18, pages 7168–7183. DOI: 10.1016/j.eswa.2013.07.039

[144] KHOSRAVI, Khashayar ; GUÉHÉNEUC, Yann-Gaël: **"A quality model for design patterns"**. In *University of Montreal, Tech. Rep* (2004)

[145] KIM, Joseph ; MUISE, Christian ; SHAH, Ankit ; AGARWAL, Shubham ; SHAH, Julie: **"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

[146] KIM, Yongho ; JUNG, Jin-Woo ; GALLAGHER, John C. ; MATSON, Eric T.: **"An adaptive goal-based model for autonomous multi-robot using HARMS and NuSMV"**. In *International Journal of Fuzzy Logic and Intelligent Systems* 16 (2016), number 2, pages 95–103

[147] KIM, Yongho ; JUNG, Jin-Woo ; MATSON, Eric T.: **"An Adaptive Task-Based Model for Autonomous Multi-Robot Using HARMS and NuSMV."**. In *FNC/MobiSPC*, 2015, pages 127–132

[148] KINNY, D ; GEORGE, M: **"Commitment and E ectiveness of Situated Agents"**. In *Proceedings of the twelfth international joint conference on artificial intelligence (IJCAI-91)*, 1991, pages 82–88

[149] KIRAN, Mariam ; RICHMOND, Paul ; HOLCOMBE, Mike ; CHIN, Lee S. ; WORTH, David ; GREENOUGH, Chris: **"FLAME: simulating large populations of agents on parallel hardware architectures"**. In *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems: volume 1* Int. Foundation for Autonomous Agents and Multiagent Systems (event), 2010, pages 1633–1636

[150] KITCHENHAM, B. ; CHARTERS, S: *Guidelines for performing Systematic Literature Reviews in Software Engineering*. 2007

[151] KITCHENHAM, Barbara A. ; BRERETON, O P. ; BUDGEN, David ; LI, Zhi: **"An Evaluation of Quality Checklist Proposals-A participant-observer case study."**. In *EASE* Volume 9, 2009, pages 167

[152] KITCHENHAM, Barbara A. ; BRERETON, Pearl ; TURNER, Mark ; NIAZI, Mahmood K. ; LINKMAN, Stephen ; PRETORIUS, Rialette ; BUDGEN, David: **"Refining the systematic literature review process—two participant-observer case studies"**. In *Empirical Software Engineering* 15 (2010), number 6, pages 618–653

[153] KOENIG, Nathan ; HOWARD, Andrew: **"Design and use paradigms for gazebo, an open-source multi-robot simulator"**. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)* Volume 3 IEEE (event), 2004, pages 2149–2154

[154] KOESTLER, A: *The Ghost in the Machine, Hutchinson*. 1967

[155] KOLLING, Andreas ; WALKER, Phillip ; CHAKRABORTY, Nilanjan ; SYCARA, Katia ; LEWIS, Michael: **"Human interaction with robot swarms: A survey"**. In *IEEE Transactions on Human-Machine Systems* 46 (2016), number 1, pages 9–26

[156] KRAVARI, Kalliopi ; BASSILIADES, Nick: **"A survey of agent platforms"**. In *Journal of Artificial Societies and Social Simulation* 18 (2015), number 1, pages 11

[157] KREBS, Friedrich ; ERNST, Andreas: **"A spatially explicit agent-based model of the diffusion of green electricity: Model setup and retrodictive validation"**. In *Advances in Social Simulation*. Springer, 2017, pages 217–230

[158] KRIZEK, Gerd C.: *Ockham's razor and the interpretations of quantum mechanics*. 2017

[159] KUCHEROV, Dmytro P. ; KUCHEROV, DP: **"Group behavior of UAVs in obstacles presence"**. In *4th Int. Conf. on Methods and Systems of Navigation and Motion Control (MSNMC)*, National aviation university, 2016, pages 51–54. DOI: 10.1109/MSNMC.2016.7783104

[160] KULESZA, Todd ; STUMPF, Simone ; BURNETT, Margaret ; KWAN, Irwin: **"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, pages 1–10

[161] KULESZA, Todd ; STUMPF, Simone ; BURNETT, Margaret ; WONG, Weng-Keen ; RICHE, Yann ; MOORE, Travis ; OBERST, Ian ; SHINSEL, Amber ; MCINTOSH, Kevin: **"Explanatory debugging: Supporting end-user debugging of machine-learned programs"**. In *IEEE Symposium on Visual Languages and Human-Centric Computing* IEEE (event), 2010, pages 41–48

[162] KULESZA, Todd ; STUMPF, Simone ; BURNETT, Margaret ; YANG, Sherry ; KWAN, Irwin ; WONG, Weng-Keen: **"Too much, too little, or just right? Ways explanations impact end users' mental models"**. In *IEEE Symposium on Visual Languages and Human Centric Computing* IEEE (event), 2013, pages 3–10

[163] KUZON, William ; URBANCHEK, Melanie ; MCCABE, Steven: **"The seven deadly sins of statistical analysis"**. In *Annals of plastic surgery* 37 (1996), pages 265–272

[164] LAIRD, John: **"The Law of Parsimony"**. In *The Monist* 29 (1919), number 3, pages 321–344. – URL http://www.jstor.org/stable/27900747

[165] LAIRD, John E.: *The Soar cognitive architecture*. MIT press, 2012

[166] LANGLEY, Pat: **"Explainable, normative, and justified agency"**. In *Proceedings of the AAAI Conference on Artificial Intelligence* Volume 33, 2019, pages 9775–9779

[167] LAW, Averill M. ; KELTON, W D. ; KELTON, W D.: *Simulation modeling and analysis*. Volume 3. McGraw-Hill New York, 2000

[168] LEE, Kuo-Chu: *Tree structured variable priority arbitration implementing a round-robin scheduling policy*. April 5 1994. – US Patent 5,301,333

[169] LEWIS, David K.: **"Causal explanation"**. In *Philosophical Papers* (1986), pages 214–240

[170] LEWIS, John ; MATSON, Eric T. ; WEI, Sherry ; MIN, Byung-Cheol: **"Implementing HARMS-based indistinguishability in ubiquitous robot organizations"**. In *Robotics and Autonomous Systems* 61 (2013), number 11, pages 1186–1192

[171] LIKERT, Rensis: **"A technique for the measurement of attitudes."**. In *Archives of psychology* (1932)

[172] LIM, Brian Y. ; DEY, Anind K.: **"Assessing demand for intelligibility in context-aware applications"**. In *Proceedings of the 11th international conference on Ubiquitous computing*, 2009, pages 195–204

[173] LIM, Brian Y. ; DEY, Anind K.: **"Design of an intelligible mobile context-aware application"**. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services* ACM (event), 2011, pages 157–166

[174] LIPPI, Giuseppe ; MATTIUZZI, Camilla: **"Biological samples transportation by drones: ready for prime time?"**. In *Annals of translational medicine* 4 (2016), number 5

[175] LIPTON, Peter: **"Contrastive explanation"**. In *Royal Institute of Philosophy Supplements* 27 (1990), pages 247–266

[176] LIPTON, Zachary C.: **"The mythos of model interpretability"**. In *Commun. ACM* 61 (2018), number 10, pages 36–43. – URL https://doi.org/10.1145/3233231. DOI: 10.1145/3233231

[177] LIU, Hui ; YIN, Qingyu ; WANG, William Y.: **"Towards explainable NLP: A generative explanation framework for text classification"**. In *arXiv preprint arXiv:1811.00196* (2018)

[178] LOEWENSTEIN, George: **"The psychology of curiosity: A review and reinterpretation."**. In *Psychological bulletin* 116 (1994), number 1, pages 75

[179] LORIG, Fabian ; DAMMENHAYN, Nils ; MÜLLER, David-Johannes ; TIMM, Ingo J.: **"Measuring and Comparing Scalability of Agent-Based Simulation Frameworks"**. In *German Conf. on Multiagent System Technologies* Springer (event), 2015, pages 42–60

[180] LUKE, Sean ; CIOFFI-REVILLA, Claudio ; PANAIT, Liviu ; SULLIVAN, Keith ; BALAN, Gabriel C.: **"MASON: A Multiagent Simulation Environment"**. In *Simulation* 81 (2005), number 7, pages 517–527

[181] MA, Xiaobai ; JIAO, Ziyuan ; WANG, Zhenkai ; PANAGOU, Dimitra: **"Decentralized prioritized motion planning for multiple autonomous UAVs in 3D polygonal obstacle environments"**. In *Int. Conf. on Unmanned Aircraft Systems*, 2016

[182] MACAL, Charles M. ; NORTH, Michael J.: **"Tutorial on agent-based modeling and simulation"**. In *Proceedings of the Winter Simulation Conference, 2005.* IEEE (event), 2005, pages 14–pp

[183] MACH, Ernst: *The science of mechanics*. Prabhat Prakashan, 1919

[184] MADHAVAN, Poornima ; WIEGMANN, Douglas A.: **"Effects of information source, pedigree, and reliability on operator interaction with decision support systems"**. In *Human Factors* 49 (2007), number 5, pages 773–785

[185] MADUMAL, Prashan ; MILLER, Tim ; SONENBERG, Liz ; VETERE, Frank: **"Explainable reinforcement learning through a causal lens"**. In *arXiv preprint arXiv:1905.10958* (2019)

[186] MALLE, Bertram F.: **"How people explain behavior: A new theoretical framework"**. In *Personality and social psychology review* 3 (1999), number 1, pages 23–48

[187] MALLE, Bertram F.: *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press, 2006

[188] MATHEWS, Sherin M.: **"Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review"**. In *Intelligent Computing-Proceedings of the Computing Conference* Springer (event), 2019, pages 1269–1292

[189] MATSON, Eric T. ; MIN, Byung-Cheol: **"M2M infrastructure to integrate humans, agents and robots into collectives"**. In *IEEE International Instrumentation and Measurement Technology Conference* IEEE (event), 2011, pages 1–6

[190] MATSON, Eric T. ; TAYLOR, Julia ; RASKIN, Victor ; MIN, Byung-Cheol ; WILSON, E C.: **"A natural language exchange model for enabling human, agent, robot and machine interaction"**. In *The 5th International Conference on Automation, Robotics and Applications* IEEE (event), 2011, pages 340–345

[191] MCEVOY, John F. ; HALL, Graham P. ; MCDONALD, Paul G.: **"Evaluation of unmanned aerial vehicle shape, flight path and camera type for waterfowl surveys: disturbance effects and species recognition"**. In *PeerJ* 4 (2016), pages e1831

[192] MCGUINNESS, Deborah L. ; VAN HARMELEN, Frank ; OTHERS: **"OWL web ontology language overview"**. In *W3C recommendation* 10 (2004), number 10, pages 2004

[193] MCLAREN, Duncan ; AGYEMAN, Julian: *Sharing Cities: A Case for Truly Smart and Sustainable Cities*. MIT press, 2015. – ISBN 9780262029728

[194] MEFTEH, Wafa ; MIGEON, Frédéric ; GLEIZES, Marie-Pierre ; GARGOURI, Faiez: **"Simulation Based Design for Adaptive Multi-agent Systems: Extension to the ADELFE Methdology"**. In *2013 Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises* IEEE (event), 2013, pages 36–38

[195] MENOUAR, Hamid ; GUVENC, Ismail ; AKKAYA, Kemal ; ULUAGAC, A S. ; KADRI, Abdullah ; TUNCER, Adem: **"UAV-enabled intelligent transportation systems**

for the smart city: Applications and challenges". In *IEEE Communications Magazine* 55 (2017), number 3, pages 22–28

[196] MICHON, J.A.: **"A critical view of driver behaviour models: What do we know, what should we do?"**. In *Human Behavior and Traffic Safety* (1985), pages 487–525

[197] MIHALY, Heder: **"From NASA to EU: the evolution of the TRL scale in Public Sector Innovation"**. In *The Innovation Journal* 22 (2017), October, pages 1–23

[198] MILLER, Tim: **"Explanation in artificial intelligence: Insights from the social sciences"**. In *Artificial Intelligence* 267 (2019), pages 1–38

[199] MILLER, Tim ; HOWE, Piers ; SONENBERG, Liz: **"Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences"**. In *arXiv preprint arXiv:1712.00547* (2017)

[200] MILLIEZ, Grégoire ; LALLEMENT, Raphaël ; FIORE, Michelangelo ; ALAMI, Rachid: **"Using human knowledge awareness to adapt collaborative plan generation, explanation and monitoring"**. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* IEEE (event), 2016, pages 43–50

[201] MIN, Byung-Cheol ; HONG, Ji-Hyeon ; MATSON, Eric T.: **"Adaptive robust control (ARC) for an altitude control of a quadrotor type UAV carrying an unknown payloads"**. In *11th International Conference on Control, Automation and Systems* IEEE (event), 2011, pages 1147–1151

[202] MITTELSTADT, Brent ; RUSSELL, Chris ; WACHTER, Sandra: **"Explaining explanations in AI"**. In *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pages 279–288

[203] MOHAMMED, Farhan ; IDRIES, Ahmed ; MOHAMED, Nader ; AL-JAROODI, Jameela ; JAWHAR, Imad: **"UAVs for smart cities: Opportunities and challenges"**. In *Int. Conf. on Unmanned Aircraft Systems (ICUAS)* IEEE (event), 2014, pages 267–273

[204] MOTLAGH, Naser H. ; TALEB, Tarik ; AROUK, Osama: **"Low-altitude unmanned aerial vehicles-based internet of things services: Comprehensive survey and future perspectives"**. In *IEEE Internet of Things Journal* 3 (2016), number 6, pages 899–922

[205] MUALLA, Yazan ; BAI, Wenshuai ; GALLAND, Stéphane ; NICOLLE, Christophe: **"Comparison of Agent-based Simulation Frameworks for Unmanned Aerial Transportation Applications"**. In *Procedia computer science* 130 (2018), number C, pages 791–796. DOI: 10.1016/j.procs.2018.04.137

[206] MUALLA, Yazan ; KAMPIK, Timotheus ; TCHAPPI, Igor H. ; NAJJAR, Amro ; GALLAND, Stéphane ; NICOLLE, Christophe: **"Explainable Agents as Static Web Pages: UAV Simulation Example"**. In *Proc. of 2nd International Workshop on eXplainable TRansparent Autonomous Agents and Multi-Agent Systems, AAMAS, Auckland, New Zealand*, 2020, pages 149–154

[207] MUALLA, Yazan ; NAJJAR, Amro ; DAOUD, Alaa ; GALLAND, Stéphane ; NICOLLE, Christophe ; YASAR, Ansar-Ul-Haque ; SHAKSHUKI, Elhadi: **"Agent-based simulation of unmanned aerial vehicles in civilian applications: A systematic literature review and research directions"**. In *Future Generation Computer Systems* 100 (2019), pages 344–364. DOI: 10.1016/j.future.2019.04.051

[208] MUALLA, Yazan ; NAJJAR, Amro ; GALLAND, Stéphane ; NICOLLE, Christophe ; HAMAN TCHAPPI, Igor ; YASAR, Ansar-Ul-Haque ; FRÄMLING, Kary: **"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 (event), 2019, pages 1649–1653

[209] MUALLA, Yazan ; NAJJAR, Amro ; KAMPIK, Timotheus ; TCHAPPI, Igor ; GALLAND, Stéphane ; NICOLLE, Christophe: **"Towards Explainability for a Civilian UAV Fleet Management using an Agent-based Approach"**. In *1st Workshop on Explainable AI in Automated Driving: A User-Centered Interaction Approach, Utrecht, Netherland. arXiv preprint arXiv:1909.10090* (2019)

[210] MUALLA., Yazan ; TCHAPPI., Igor H. ; NAJJAR., Amro ; KAMPIK., Timotheus ; GALLAND., Stéphane ; NICOLLE., Christophe: **"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 (event), SciTePress, 2020, pages 378–385. – ISBN 978-989-758-395-7. DOI: 10.5220/0009382903780385

[211] NAHAVANDI, Saeid: **"Trusted autonomy between humans and robots: toward human-on-the-loop in robotics and autonomous systems"**. In *IEEE Systems, Man, and Cybernetics Magazine* 3 (2017), number 1, pages 10–17

[212] NARAYANAN, Menaka ; CHEN, Emily ; HE, Jeffrey ; KIM, Been ; GERSHMAN, Sam ; DOSHI-VELEZ, Finale: **"How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation"**. In *arXiv preprint arXiv:1802.00682* (2018)

[213] NEERINCX, Mark A. ; VUGHT, Willeke van ; HENKEMANS, Olivier B. ; OLEARI, Elettra ; BROEKENS, Joost ; PETERS, Rifca ; KAPTEIN, Frank ; DEMIRIS, Yiannis ;

KIEFER, Bernd ; FUMAGALLI, Diego ; OTHERS: **"Socio-Cognitive Engineering of a Robotic Partner for Child's Diabetes Self-Management"**. In *Frontiers in Robotics and AI* 6 (2019)

[214] NEERINCX, Mark A. ; WAA, Jasper van der ; KAPTEIN, Frank ; DIGGELEN, Jurriaan van: **"Using perceptual and cognitive explanations for enhanced human-agent team performance"**. In *International Conference on Engineering Psychology and Cognitive Ergonomics* Springer (event), 2018, pages 204–214

[215] NEGAHBAN, Ashkan ; YILMAZ, Levent: **"Agent-based simulation applications in marketing research: an integrated review"**. In *Journal of Simulation* 8 (2014), number 2, pages 129–142

[216] NIKOLAI, Cynthia ; MADEY, Gregory: **"Tools of the trade: A survey of various agent based modeling platforms"**. In *Journal of Artificial Societies and Social Simulation* 12 (2009), number 2, pages 2

[217] NORLING, Emma: **"Folk psychology for human modelling: Extending the BDI paradigm"**. In *3rd Int. Joint Conf. on Autonomous Agents and Multiagent Systems-Volume 1* IEEE Computer Society (event), 2004, pages 202–209

[218] NORMAN, Geoff: **"Likert scales, levels of measurement and the "laws" of statistics"**. In *Advances in health sciences education* 15 (2010), number 5, pages 625–632

[219] OBDRŽÁLEK, Zbyněk: **"Mobile agents in multi-agent UAV/UGV system"**. In *Military Technologies (ICMT), 2017 Int. Conf. on* IEEE (event), 2017, pages 753–759

[220] ÖREN, Tuncer: **"A critical review of definitions and about 400 types of modeling and simulation"**. In *SCS M&S Magazine* 2 (2011), number 3, pages 142–151

[221] ÖREN, Tuncer: **"The many facets of simulation through a collection of about 100 definitions"**. In *SCS M&S Magazine* 2 (2011), number 2, pages 82–92

[222] OSIPOW, Samuel H.: **"Theories of Career Development. A Comparison of the Theories."**. (1968)

[223] OZPINECI, Burak ; TOLBERT, Leon M.: **"Simulink implementation of induction machine model-a modular approach"**. In *IEEE International Conference of Electric Machines and Drives* Volume 2 Ieee (event), 2003, pages 728–734

[224] PARUNAK, H.D.: **"Making swarming happen"**. In *Conf. on Swarming and Network Enabled Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance (C4ISR)*. McLean, Virginia, USA, January 2003

[225] PASCARELLA, Domenico ; VENTICINQUE, Salvatore ; AVERSA, Rocco: **"Agent-based design for UAV mission planning"**. In *8th Int. Conf. on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)* IEEE (event), 2013, pages 76–83

[226] PATEL, Ankit ; DEGESYS, Julius ; NAGPAL, Radhika: **"Desynchronization: The theory of self-organizing algorithms for round-robin scheduling"**. In *First International Conference on Self-Adaptive and Self-Organizing Systems (SASO 2007)* IEEE (event), 2007, pages 87–96

[227] PAULWEBER, Michael ; LEBERT, Klaus: *Powertrain Instrumentation and Test Systems: Development - Hybridization - Electrification*. Springer, 2016 DOI: 10.1007/978-3-319-32135-6

[228] PECHOUCEK, Michal ; JAKOB, Michal ; NOVÁK, Peter: **"Towards Simulation-Aided Design of Multi-Agent Systems"**. In *Programming Multi-Agent Systems - 8th Int. Workshop, ProMAS 2010, Toronto, ON, Canada, May 11, 2010. Revised Selected Papers*, 2010, pages 3–21. DOI: 10.1007/978-3-642-28939-2_1

[229] PĚCHOUČEK, Michal ; MAŘÍK, Vladimír: **"Industrial deployment of multi-agent technologies: review and selected case studies"**. In *Autonomous agents and multi-agent systems* 17 (2008), number 3, pages 397–431

[230] PENG, Zhi-hong ; WU, Jin-ping ; CHEN, Jie: **"Three-dimensional multi-constraint route planning of unmanned aerial vehicle low-altitude penetration based on coevolutionary multi-agent genetic algorithm"**. In *Journal of Central South University of Technology* 18 (2011), number 5, pages 1502

[231] PERIS-ORTIZ, Marta ; BENNETT, Dag R. ; PÉREZ-BUSTAMANTE YÁBAR, Diana: *Sustainable Smart Cities: Creating Spaces for Technological, Social and Business Development*. Springer, 2016. – ISBN 9783319408958

[232] PHILLIPS, P J. ; HAHN, Carina A. ; FONTANA, Peter C. ; BRONIATOWSKI, David A. ; PRZYBOCKI, Mark A.: **"Four Principles of Explainable Artificial Intelligence (Draft)"**. (2020)

[233] PREECE, Alun: **"Asking 'Why'in AI: Explainability of intelligent systems–perspectives and challenges"**. In *Intelligent Systems in Accounting, Finance and Management* 25 (2018), number 2, pages 63–72

[234] PURI, Anuj: **"A survey of unmanned aerial vehicles (UAV) for traffic surveillance"**. In *Department of computer science and engineering, University of South Florida* (2005), pages 1–29

[235] QUIJANO-SANCHEZ, Lara ; SAUER, Christian ; RECIO-GARCIA, Juan A. ; DIAZ-AGUDO, Belen: **"Make it personal: a social explanation system applied to**

group recommendations". In *Expert Systems with Applications* 76 (2017), pages 36–48

[236] RAILSBACK, Steven F. ; LYTINEN, Steven L. ; JACKSON, Stephen K.: **"Agent-based simulation platforms: Review and development recommendations"**. In *Simulation* 82 (2006), number 9, pages 609–623

[237] RAJAGOPALAN, Rajeswari P. ; KRISHNA, Rahul: **"Drones: Guidelines, regulations, and policy gaps in India"**. In *Occasional Papers* (2018)

[238] RAO, Anand S. ; GEORGEFF, Michael P. ; OTHERS: **"BDI agents: from theory to practice."**. In *ICMAS* Volume 95, 1995, pages 312–319

[239] RAS, Gabriëlle ; GERVEN, Marcel van ; HASELAGER, Pim: **"Explanation methods in deep learning: Users, values, concerns and challenges"**. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018, pages 19–36

[240] RASMUSSEN, Carl E. ; GHAHRAMANI, Zoubin: **"Occam's razor"**. In *Advances in neural information processing systems*, 2001, pages 294–300

[241] RATHI, Shubham: **"Generating counterfactual and contrastive explanations using SHAP"**. In *arXiv preprint arXiv:1906.09293* (2019)

[242] RODIĆ, Aleksandar ; MESTER, Gyula: **"Modeling and simulation of quad-rotor dynamics and spatial navigation"**. In *Intelligent Systems and Informatics (SISY), 2011 IEEE 9th Int. Symposium on* IEEE (event), 2011, pages 23–28

[243] ROLLO, Milan ; SELECKỲ, Martin ; LOSIEWICZ, Paul ; READE, John ; MAIDA, Nicholas: **"Framework for incremental development of complex unmanned aircraft systems"**. In *Integrated Communication, Navigation, and Surveillance Conf. (ICNS), 2015* IEEE (event), 2015, pages 1–14

[244] ROSENFELD, Avi ; RICHARDSON, Ariella: **"Explainability in human–agent systems"**. In *Autonomous Agents and Multi-Agent Systems* (2019), pages 1–33

[245] ROUSSET, Alban ; HERRMANN, Bénédicte ; LANG, Christophe ; PHILIPPE, Laurent: **"A survey on parallel and distributed multi-agent systems"**. In *Padabs 2nd Workshop on Parallel and Distributed Agent-Based Simulations, in conjunction with Euro-Par 2014* Volume 8805 Springer (event), 2014, pages 371–382

[246] SADO, Fatai ; LOO, Chu K. ; KERZEL, Matthias ; WERMTER, Stefan: **"Explainable Goal-Driven Agents and Robots–A Comprehensive Review and New Framework"**. In *arXiv preprint arXiv:2004.09705* (2020)

[247] SAMEK, Wojciech ; WIEGAND, Thomas ; MÜLLER, Klaus-Robert: **"Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models"**. In *arXiv preprint arXiv:1708.08296* (2017)

[248] SAMPEDRO, Carlos ; BAVLE, Hriday ; SANCHEZ LOPEZ, Jose L. ; SUAREZ FERNANDEZ, Ramon ; RODRIGUEZ RAMOS, Alejandro ; MOLINA, Martin ; CAMPOY CERVERA, Pascual: **"A flexible and dynamic mission planning architecture for uav swarm coordination"**. In *Proc. of 2016 Int. Conf. on Unmanned Aircraft Systems (ICUAS)* ETSI_Informatica (event), 2016

[249] SCHATTEN, Markus: **"Multi-agent based Traffic Control of Autonomous Unmanned Aerial Vehicles"** / Artificial Intelligence Laboratory, University of Zagreb. 2015. – Research Report

[250] SCHRAAGEN, Jan M. ; CHIPMAN, Susan F. ; SHALIN, Valerie L.: *Cognitive task analysis*. Psychology Press, 2000

[251] SCHUT, Martijn ; WOOLDRIDGE, Michael: **"The control of reasoning in resource-bounded agents"**. In *The Knowledge Engineering Review* 16 (2001), number 3, pages 215

[252] SCHUT, Martijn ; WOOLDRIDGE, Michael ; PARSONS, Simon: **"The theory and practice of intention reconsideration"**. In *Journal of Experimental & Theoretical Artificial Intelligence* 16 (2004), number 4, pages 261–293

[253] SCOTT, Judy E. ; SCOTT, Carlton H.: **"Models for Drone Delivery of Medications and Other Healthcare Items"**. In *IJHISI* 13 (2018), number 3, pages 20–34. DOI: 10.4018/IJHISI.2018070102

[254] SEMSCH, Eduard ; JAKOB, Michal ; PAVLÍCEK, Dusan ; PECHOUCEK, Michal: **"Autonomous UAV Surveillance in Complex Urban Environments"**. In *Proc. of 2009 IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology, IAT 2009, Milan, Italy, 15-18 September 2009*, 2009, pages 82–85. DOI: 10.1109/WI-IAT.2009.132

[255] SERENKO, Alexander ; DETLOR, Brian ; OTHERS: **"Agent toolkits: A general overview of the market and an assessment of instructor satisfaction with utilizing toolkits in the classroom"**. In *Research and working paper series of the Michael G. DeGroote School of Business* (2002), July, number 455, pages 43

[256] SHAH, Shital ; DEY, Debadeepta ; LOVETT, Chris ; KAPOOR, Ashish: **"AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles"**. In *Field and Service Robotics*, 2017

[257] SHAHEEN, Susan ; STOCKER, Adam ; MUNDLER, Marie: **"Online and app-based carpooling in France: Analyzing users and practices—A study of BlaBlaCar"**. In *Disrupting Mobility*. Springer, 2017, pages 181–196

[258] SHAKHATREH, Hazim ; SAWALMEH, Ahmad ; AL-FUQAHA, Ala ; DOU, Zuochao ; ALMAITA, Eyad ; KHALIL, Issa ; OTHMAN, Noor S. ; KHREISHAH, Abdallah ; GUIZANI, Mohsen: **"Unmanned Aerial Vehicles: A Survey on Civil Applications and Key Research Challenges"**. In *arXiv preprint arXiv:1805.00881* (2018)

[259] SHARMA, Deepak ; BHONDEKAR, Amol P. ; SHUKLA, AK ; GHANSHYAM, C: **"A review on technological advancements in crowd management"**. In *Journal of Ambient Intelligence and Humanized Computing* (2016), pages 1–11

[260] SHOHAM, Yoav: **"Agent-oriented programming"**. In *Artificial intelligence* 60 (1993), number 1, pages 51–92

[261] SILVA, Luis C B. da ; BERNARDO, Ricardo M. ; OLIVEIRA, Hugo A. de ; ROSA, Paulo F.: **"Multi-UAV agent-based coordination for persistent surveillance with dynamic priorities"**. In *Int. Conf. on Military Technologies (ICMT)* IEEE (event), 2017, pages 765–771

[262] ŠIŠLÁK, David ; ROLLO, Milan ; PĚCHOUČEK, Michal: **"A-globe: Agent platform with inaccessibility and mobility support"**. In *Int. Workshop on Cooperative Information Agents* Springer (event), 2004, pages 199–214

[263] ŠIŠLÁK, David ; VOLF, Přemysl ; KOPŘIVA, Štěpán ; PĚCHOUČEK, Michal: **"Agent-fly: NAS-wide simulation framework integrating algorithms for automated collision avoidance"**. In *Integrated Communications, Navigation and Surveilance Conf. (ICNS), 2011* IEEE (event), 2011, pages F7/1–F7/11

[264] SISLAK, David ; VOLF, Premysl ; PECHOUCEK, Michal ; CANNON, Christopher T. ; NGUYEN, Duc N. ; REGLI, William C.: **"Multi-Agent Simulation of En-Route Human Air-Traffic Controller"**. In *the Twenty-Fourth Innovative Appications of Artificial Intelligence Conference*. Toronto, Canada : AAAI Press, 2012, pages 2323–2328. – ISBN 978-1-57735-568-7

[265] STEFANI, Antonia ; STAVRINOUDIS, Dimitris ; XENOS, Michalis: **"Experimental based tool calibration used for assessing the quality of e-commerce systems"**. In *e-Business and Telecommunication Networks*. Springer, 2006, pages 100–106

[266] STENGER, A ; FERNANDO, B ; HENI, M: *Autonomous Mission Planning for UAVs: A Cognitive Approach*. Citeseer, 2013

[267] STUMPF, Simone ; RAJARAM, Vidya ; LI, Lida ; BURNETT, Margaret ; DIETTERICH, Thomas ; SULLIVAN, Erin ; DRUMMOND, Russell ; HERLOCKER, Jonathan: **"Toward harnessing user feedback for machine learning"**. In *Proceedings of the 12th international conference on Intelligent user interfaces*, 2007, pages 82–91

[268] SUKKERD, Roykrong ; SIMMONS, Reid ; GARLAN, David: **"Towards Explainable Multi-Objective Probabilistic Planning"**. In *Proceedings of the 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS\'18)*, 2018

[269] SULLIVAN, Gail M. ; ARTINO JR, Anthony R.: **"Analyzing and interpreting data from Likert-type scales"**. In *Journal of graduate medical education* 5 (2013), number 4, pages 541–542

[270] SUN, Ron ; MERRILL, Edward ; PETERSON, Todd: **"A bottom-up model of skill learning"**. In *Proc. of 20th cognitive science society conference*, 1998, pages 1037–1042

[271] SUTTON, Andrew ; FIDAN, Barış ; WALLE, Dirk Van der: **"Hierarchical uav formation control for cooperative surveillance"**. In *IFAC Proc. Volumes* 41 (2008), number 2, pages 12087–12092

[272] SWARTOUT, William ; PARIS, Cecile ; MOORE, Johanna: **"Explanations in knowledge systems: Design for explainable expert systems"**. In *IEEE Expert* 6 (1991), number 3, pages 58–64

[273] SWELLER, John: **"Cognitive load theory"**. In *Psychology of learning and motivation* Volume 55. 2011, pages 37–76

[274] SYCARA, Katia P.: **"Multiagent systems"**. In *AI magazine* 19 (1998), number 2, pages 79

[275] SZEGEDY, Christian ; ZAREMBA, Wojciech ; SUTSKEVER, Ilya ; BRUNA, Joan ; ERHAN, Dumitru ; GOODFELLOW, Ian ; FERGUS, Rob: **"Intriguing properties of neural networks"**. In *arXiv preprint arXiv:1312.6199* (2013)

[276] TAILLANDIER, Patrick ; BOURGAIS, Mathieu ; DROGOUL, Alexis ; VERCOUTER, Laurent: **"Using parallel computing to improve the scalability of models with BDI agents"**. In *Social Simulation Conference* Springer (event), 2017, pages 37–47

[277] TAILLANDIER, Patrick ; BOURGAIS, Mathieu ; DROGOUL, Alexis ; VERCOUTER, Laurent: **"the Scalability of Models with BDI Agents"**. In *Social Simulation for a Digital Society: Applications and Innovations in Computational Social Science* (2019), pages 37

[278] TANIA, Lombrozo: **"The structure and function of explanations"**. In *Trends in Cognitive Sciences* 10 (2006), number 10, pages 464–470

[279] TAYLOR, Glenn ; KNUDSEN, Keith ; HOLT, Lisa S.: **"Explaining agent behavior"**. In *Proceedings of the 15th Conference on Behavior Representation in Modeling and Simulation (BRIMS06)*, 2006

[280] TEAL GROUP: **"World Unmanned Aerial Vehicle Systems – 2018 Market Profile and Forecast"** / Teal Group Corporation. 07 2018. – Research Report

[281] THE UNITED NATIONS ORGANIZATION: *68% of the world population projected to live in urban areas by 2050, says UN*. 05 2018. – https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html. Accessed on 2019-02-20

[282] THE WORLD BANK: *Rural population (% of total population)*. 2018. – https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS. Accessed on 2019-02-20

[283] THEODOROPOULOS, Georgios ; MINSON, Rob ; EWALD, Roland ; LEES, Michael: **"Simulation engines for multi-agent systems"**. In *Multi-Agent Systems: Simulation and Applications", edited by Adelinde M. Uhrmacher and Danny Weyns (Editors), Publisher: Taylor and Fran-cis. ISBN* 1779537239 (2009), pages 77–108

[284] THORBURN, William M.: **"The myth of Occam's razor"**. In *Mind* 27 (1918), number 107, pages 345–353

[285] TOBIAS, Robert ; HOFMANN, Carole: **"Evaluation of free Java-libraries for social-scientific agent based simulation"**. In *Journal of Artificial Societies and Social Simulation* 7 (2004), number 1

[286] TORENS, Christoph ; ADOLF, Florian-M. ; GOORMANN, Lukas: **"Certification and Software Verification Considerations for Autonomous Unmanned Aircraft"**. In *Journal of Aerospace Computing, Information and Communication* (2014)

[287] TREUIL, J. P. ; A., Drogoul ; ZUCKER, J. D.: *Modélisation et simulation à base d'agents*. Dunod, 2008

[288] TSIAKAS, Konstantinos ; BARAKOVA, Emilia ; KHAN, Javed V. ; MARKOPOULOS, Panos: **"BrainHood: towards an explainable recommendation system for self-regulated cognitive training in children"**. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 2020, pages 1–6

[289] TULLIO, Joe ; DEY, Anind K. ; CHALECKI, Jason ; FOGARTY, James: **"How it works: a field study of non-technical users interacting with an intelligent system"**. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007, pages 31–40

[290] UNESCO: *How much does your country invest in R&D?* 2018. – URL http://uis.unesco.org/apps/visualisations/research-and-development-spending/

[291] UNION, Innovation: *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*. Brussels, 2014

[292] VASILIJEVIC, A ; CALADO, Pedro ; LOPEZ-CASTEJON, Francisco ; HAYES, Dan ; STILINOVIC, N ; NAD, D ; MANDIC, F ; DIAS, P ; GOMES, J ; MOLINA, JC ; OTHERS: **"Heterogeneous robotic system for underwater oil spill survey"**. In *OCEANS 2015-Genova* IEEE (event), 2015, pages 1–7

[293] VECHT, Bob van der ; DIGGELEN, Jurriaan van ; PEETERS, Marieke ; BARNHOORN, Jonathan ; WAA, Jasper van der: **"SAIL: a social artificial intelligence layer for human-machine teaming"**. In *Int. Conf. on Practical Applications of Agents and Multi-Agent Systems* Springer (event), 2018, pages 262–274

[294] VELOSO, Ruben ; KOKKINOGENIS, Zafeiris ; PASSOS, Lucio S. ; OLIVEIRA, Gustavo ; ROSSETTI, Rosaldo J. ; GABRIEL, Joaquim: **"A platform for the design, simulation and development of quadcopter multi-agent systems"**. In *9th Iberian Conf. on Information Systems and Technologies (CISTI)* IEEE (event), 2014, pages 1–6

[295] VERSTAEVEL, Nicolas ; RÉGIS, Christine ; GLEIZES, Marie-Pierre ; ROBERT, Fabrice: **"Principles and experimentations of self-organizing embedded agents allowing learning from demonstration in ambient robotics"**. In *Future Generation Computer Systems* 64 (2016), pages 78–87

[296] VIKHAR, P. A.: **"Evolutionary algorithms: A critical review and its future prospects"**. In *the 2016 Int. Conf. on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*. 2016, pages 261–265

[297] VOGELTANZ, Tomáš: **"A survey of free software for the design, analysis, modelling, and simulation of an unmanned aerial vehicle"**. In *Archives of Computational Methods in Engineering* 23 (2016), number 3, pages 449–514

[298] VOIGT, Paul ; BUSSCHE, Axel Von dem: **"The EU General Data Protection Regulation (GDPR)"**. In *A Practical Guide, 1st Ed., Cham: Springer International Publishing* (2017)

[299] VOLF, Premysl ; SISLÁK, David ; PECHOUCEK, Michal: **"Large-Scale High-Fidelity Agent-Based Simulation in Air Traffic Domain"**. In *Cybernetics and Systems* 42 (2011), number 7, pages 502–525. DOI: 10.1080/01969722.2011.610270

[300] WAGONER, Amy R. ; MATSON, Eric T.: **"A Robust Human-Robot Communication System Using Natural Language for HARMS."**. In *FNC/MobiSPC*, 2015, pages 119–126

[301] WALLE, Dirk Van der ; FIDAN, Baris ; SUTTON, Andrew ; YU, Changbin ; ANDERSON, Brian D.: **"Non-hierarchical UAV formation control for surveillance tasks"**. In *American Control Conference, 2008* IEEE (event), 2008, pages 777–782

[302] WANG, Danding ; YANG, Qian ; ABDUL, Ashraf ; LIM, Brian Y.: **"Designing theory-driven user-centric explainable AI"**. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pages 1–15

[303] WANG, Ning ; PYNADATH, David V. ; HILL, Susan G.: **"The impact of POMDP-generated explanations on trust and performance in human-robot teams"**. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems* International Foundation for Autonomous Agents and Multiagent Systems (event), 2016, pages 997–1005

[304] WEBSTER, Matthew P. ; CAMERON, Neil ; FISHER, Michael ; JUMP, Mike: **"Generating Certification Evidence for Autonomous Unmanned Aircraft Using Model Checking and Simulation"**. In *J. Aerospace Inf. Sys.* 11 (2014), number 5, pages 258–279. DOI: 10.2514/1.I010096

[305] WEBSTER, Matthew P. ; FISHER, Michael ; CAMERON, Neil ; JUMP, Mike: **"Formal Methods for the Certification of Autonomous Unmanned Aircraft Systems"**. In *Computer Safety, Reliability, and Security - 30th Int. Conference, SAFECOMP 2011, Naples, Italy, September 19-22*, 2011, pages 228–242. DOI: 10.1007/978-3-642-24270-0_17

[306] WEI, Yi ; BLAKE, M. B. ; MADEY, Gregory R.: **"An Operation-Time Simulation Framework for UAV Swarm Configuration and Mission Planning"**. In *Proc. of Int. Conf. on Computational Science, ICCS 2013, Barcelona, Spain, 5-7 June, 2013*, 2013, pages 1949–1958. DOI: 10.1016/j.procs.2013.05.364

[307] WEI, Yi ; MADEY, Gregory R. ; BLAKE, M. B.: **"Agent-based simulation for UAV swarm mission planning and execution"**. In *Proc. of Agent-Directed Simulation Symposium, part of the 2013 Spring Simulation Multiconference, SpringSim '13, San Diego, CA, USA, April 07 - 10, 2013*, 2013, pages 2

[308] WEIÍ, C.: **"V2X communication in Europe - from research projects towards standardization and field testing of vehicle communication technology"**. In *Comput. Netw.* 55 (2011), October, number 14, pages 3103–3119

[309] WEISS, Gerhard: *Multiagent Systems*. Boston, MA, USA : The MIT Press, 2013 (Intelligent Robotics and Autonomous Agents). – ISBN 9780262018890

[310] WERNER, David ; CRUZ, Christophe ; NICOLLE, Christophe: **"Ontology-based recommender system of economic articles"**. In *arXiv preprint arXiv:1301.4781* (2013)

[311] WEYNS, Danny ; OMICINI, Andrea ; ODELL, James: **"Environment as a First-class Abstraction in Multi-Agent Systems"**. In *Autonomous Agents and Multi-Agent Systems* 14 (2007), February, number 1, pages 5–30. – Special Issue on Environments for Multi-agent Systems. – ISSN 1387-2532

[312] WILD, Graham ; MURRAY, John ; BAXTER, Glenn: **"Exploring Civil Drone Accidents and Incidents to Help Prevent Potential Air Disasters"**. In *Aerospace* 3 (2016), number 3, pages 22

[313] WILENSKY, Uri ; EVANSTON, I: **"NetLogo: Center for connected learning and computer-based modeling"**. In *Northwestern University, Evanston, IL* 4952 (1999)

[314] WINIKOFF, Michael: **"Debugging agent programs with why? questions"**. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017, pages 251–259

[315] WITTGENSTEIN, Ludwig: *Tractatus logico-philosophicus*. Gallimard, 2001. – 121 pages. – ISBN 978-2-07-075864-7

[316] WOOLDRIDGE, Michael: **"Intelligent Agents"**. In *Multiagent systems: a modern approach to distributed artificial intelligence*. MIT press, 1999, pages 27–76

[317] WOOLDRIDGE, Michael: *An introduction to multiagent systems*. John Wiley & Sons, 2009

[318] WOOLDRIDGE, Michael ; JENNINGS, Nicholas R.: **"Intelligent agents: Theory and practice"**. In *The knowledge engineering review* 10 (1995), number 2, pages 115–152

[319] YUAN, Yifei ; WANG, Zhenrui ; LI, Mingyang ; SON, Young-Jun ; LIU, Jian: **"DDDAS-based information-aggregation for crowd dynamics modeling with UAVs and UGVs"**. In *Frontiers in Robotics and AI* 2 (2015), pages 8

[320] ZEIGLER, Bernard ; MUZY, Alexandre ; YILMAZ, Levent: *Artificial Intelligence in Modeling and Simulation*. pages 344–368. In MEYERS, Robert A. (editors): *Encyclopedia of Complexity and Systems Science*. New York, NY : Springer New York, 2009. – URL https://doi.org/10.1007/978-0-387-30440-3_24. – ISBN 978-0-387-30440-3. DOI: 10.1007/978-0-387-30440-3_24

[321] ZEIGLER, Bernard P. ; KIM, Tag G. ; PRAEHOFER, Herbert: *Theory of modeling and simulation*. Academic press, 2000

[322] ZHANG, Quan-shi ; ZHU, Song-Chun: **"Visual interpretability for deep learning: a survey"**. In *Frontiers of Information Technology & Electronic Engineering* 19 (2018), number 1, pages 27–39

[323] ZHAO, Zhongliang ; BRAUN, Torsten: **"Topology Control and Mobility Strategy for UAV Ad-hoc Networks: A Survey"**. In *Joint ERCIM eMobility and MobiSense Workshop* Citeseer (event), 2012, pages 27–32

[324] ZHOU, Zhi ; CHAN, Wai Kin V. ; CHOW, Joe H.: **"Agent-based simulation of electricity markets: a survey of tools"**. In *Artificial Intelligence Review* 28 (2007), number 4, pages 305–342

[325] ZHU, Xueping ; LIU, Zhengchun ; YANG, Jun: **"Model of Collaborative UAV Swarm Toward Coordination and Control Mechanisms Study"**. In *Proc. of Int. Conf. on Computational Science, ICCS, Computational Science at the Gates of Nature, Reykjavík, Iceland, 1-3 June*, 2015, pages 493–502. DOI: 10.1016/j.procs.2015.05.274

[326] ZOU, Xueyi ; ALEXANDER, Rob ; MCDERMID, John: **"Testing Method for Multi-UAV Conflict Resolution Using Agent-Based Simulation and Multi-Objective Search"**. In *J. Aerospace Inf. Sys.* 13 (2016), number 5, pages 191–203. DOI: 10.2514/1.I010412

List of Figures

List of Tables

List of Definitions

VI

APPENDIXES

A

The Rest of Systematic Literature Review Results

This appendix includes some extra content from the Systematic Literature Review (SLR) presented in Chapter 4 and is based on the journal article:

Yazan Mualla, Amro Najjar, Alaa Daoud, Stéphane Galland, Christophe Nicolle, Ansar-Ul-Haque Yasar, and Elhadi Shakshuki, *Agent-Based Simulation of Unmanned Aerial Vehicles in Civilian Applications: A Systematic Literature Review and Research Directions*, International Journal of Future Generation Computer Systems, Elsevier, vol. 100, pp. 344-364 (2019). DOI: 10.1016/j.future.2019.04.051.

It includes information about the geographical distributions of papers before and after the fine-grained inclusion/exclusion step. Figure A.1 plots the geographical distributions of papers before the fine-grained inclusion/exclusion step. The number of papers published by USA researchers is the highest worldwide. The geographical distribution of the papers could be partly explained by the investment rate in Research & Development (R&D) in each country [290], illustrated in Figure A.2. The notable exception is China, which invests 2% of its Gross Domestic Product (GDP), *i.e.* $US\$370,589.7$M, but has a number of papers equal to France (2.3% of GDP, US$\$60,781.6$M). Another notable point is the Czech Rep., which has 11 papers with an average R&D investment (2% of GDP, US$\$6,719$M). All the authors from Czech Rep. are collaborating with partners within USA-funded projects. This fact may explain the high number of publications for this country.

From Figure A.1, it is interesting to note that, even if civilian Unmanned Aerial Vehicles (UAVs) regulations in the USA are less restrictive than in Europe (*see* Section 10.2.3.2), the number of papers over Europe is two times higher than the number of papers in North America. This can be attributed to the European Union's (EU) research policy that enforces funding on breaking technologies, such as UAVs. Figure A.3 plots the number of papers per year since 2008 after the coarse-grained exclusion/inclusion step. The number of papers grows with a slop of 0.6364.

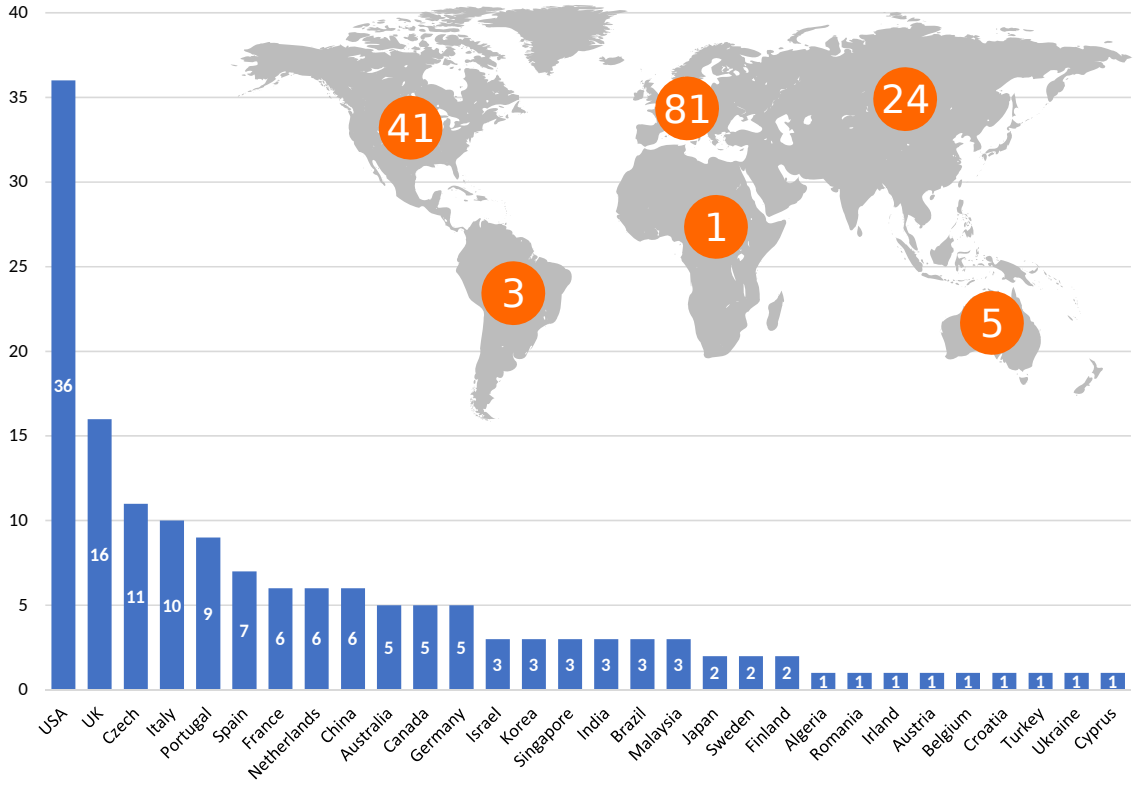


Figure A.1: The geographical distribution of papers after the coarse-grained exclusion/inclusion step

Second, this appendix also includes the 5 SLR Questions (SLRQs) that are not included in Chapter 4. These SLRQs are SLRQ4, SLRQ5, SLRQ6, SLRQ7, and SLRQ8 (refer to Section 4.2.1 for more details).

A.1/ Demographic Data (SLRQ4)

To understand the evolution of UAV simulations in Multi-agent System (MAS) in the last decade (stated as a question by SLRQ4), Figure A.4 plots the number of papers per year after the fine-grained exclusion/inclusion step. Despite a decrease in the number of papers in 2009 and 2012 (2018 should not be considered since this review was conducted in August 2018), it appears from the figure that there is a stable growth in the numbers of papers, with a slop of 0.2727. Furthermore, comparing the results of this figure with those of Figure A.3 confirms this observation since in Figure A.3, the number of papers witnesses a roughly stable growth between 2008 and 2016.

To understand the geographic distributions of the main contributors in the studied domain, Figure A.5 plots the number of papers per country after the fine-grained exclusion/inclusion step. Compared with Figure A.1, The number of papers published by USA

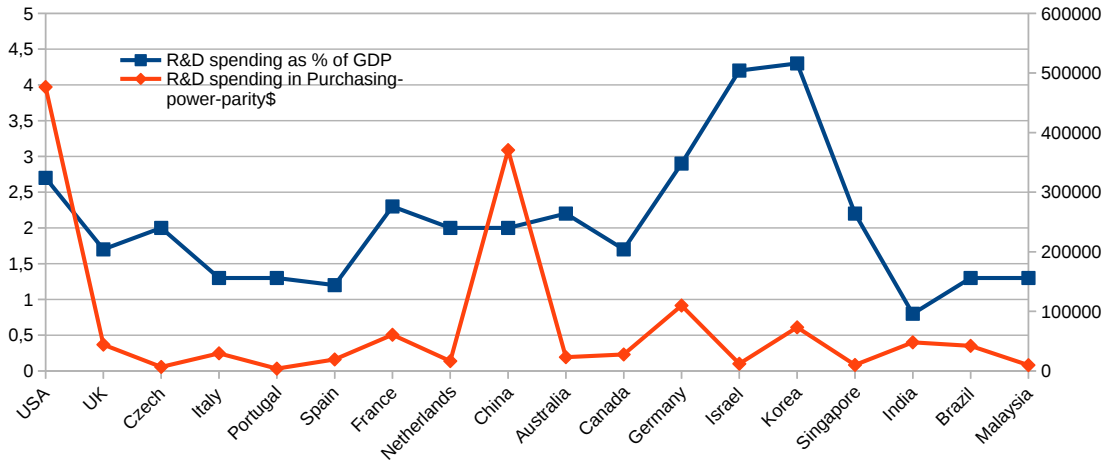


Figure A.2: The investments in R&D for the 18 most publishing countries [290] after the coarse-grained exclusion/inclusion step

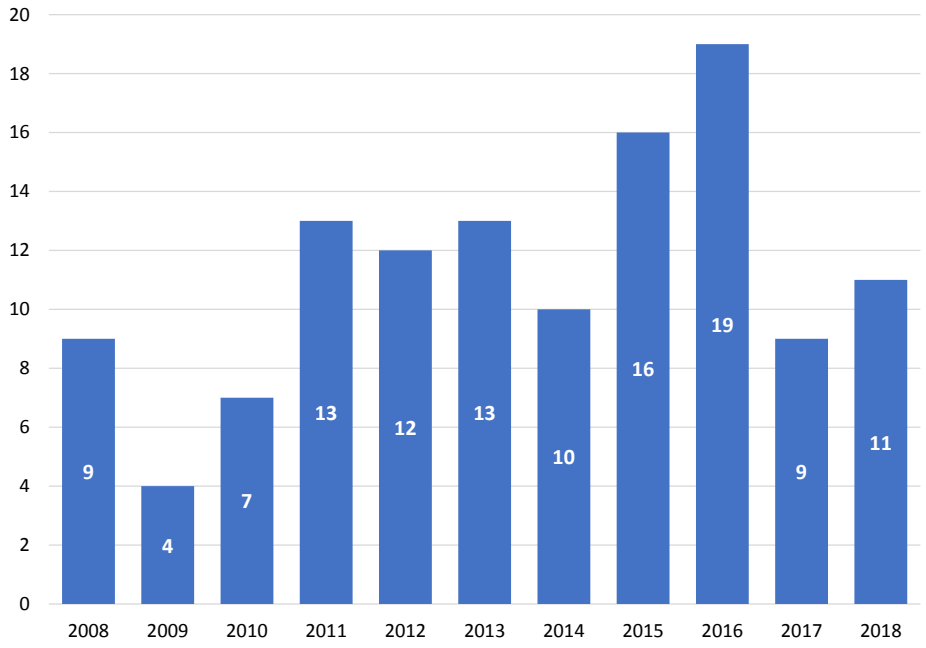


Figure A.3: The number of papers per year after the coarse-grained exclusion/inclusion step

researchers is still the highest worldwide. As previously noticed, researchers from the Czech Rep. are collaborating with partners within USA-funding projects. This fact may explain the high number of publications for this country compared to its R&D investment, illustrated in Figure A.6. In this figure, it is interesting to note that, even if civilian UAV-related regulations in the USA are less restrictive than in Europe, the number of papers over Europe (32 papers) is three times higher than the number of papers in North America (9 papers). We explain this by the fact that the EU research policy enforces funding on breaking technologies, such as UAVs. Additionally, with a relatively lower R&D investment

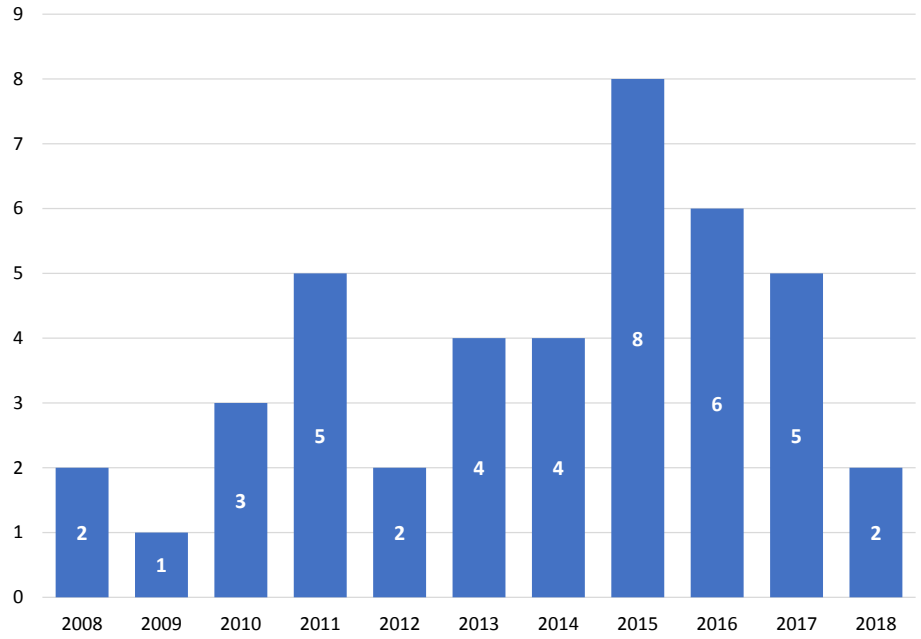


Figure A.4: The number of papers per year after the fine-grained exclusion/inclusion step

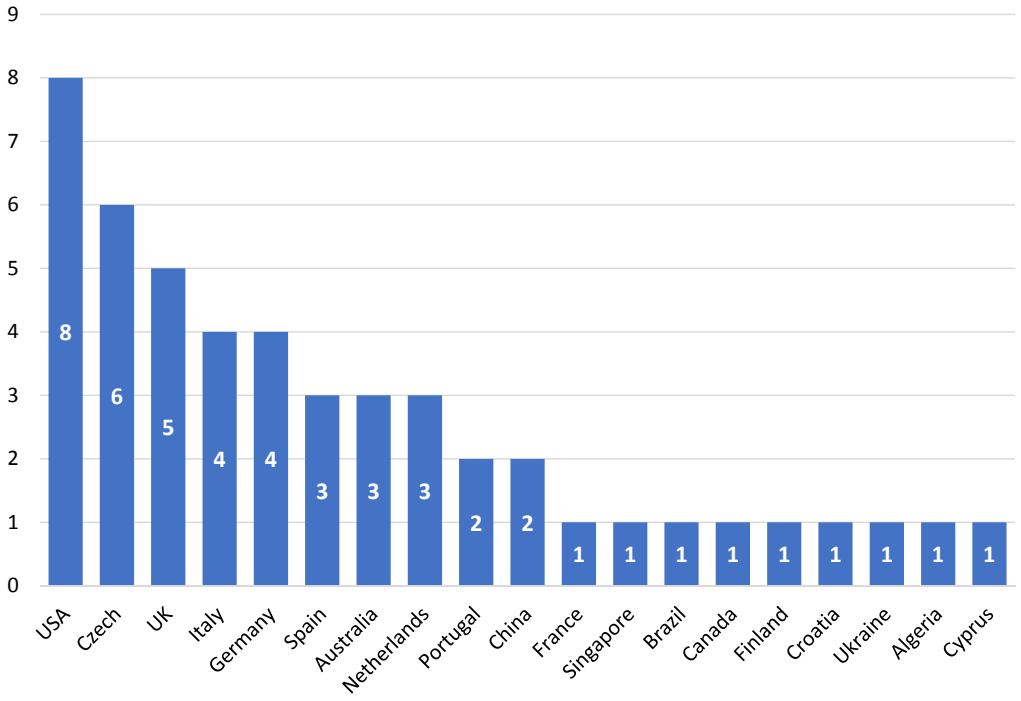


Figure A.5: The geographical distribution of papers after the fine-grained exclusion/inclusion step

rate of EU countries (Figure A.6), researchers from these countries have fewer opportunities for funding UAV deployment in real fields. This pushes them to resort to simulation environments for validating the UAV behaviors before any deployment.

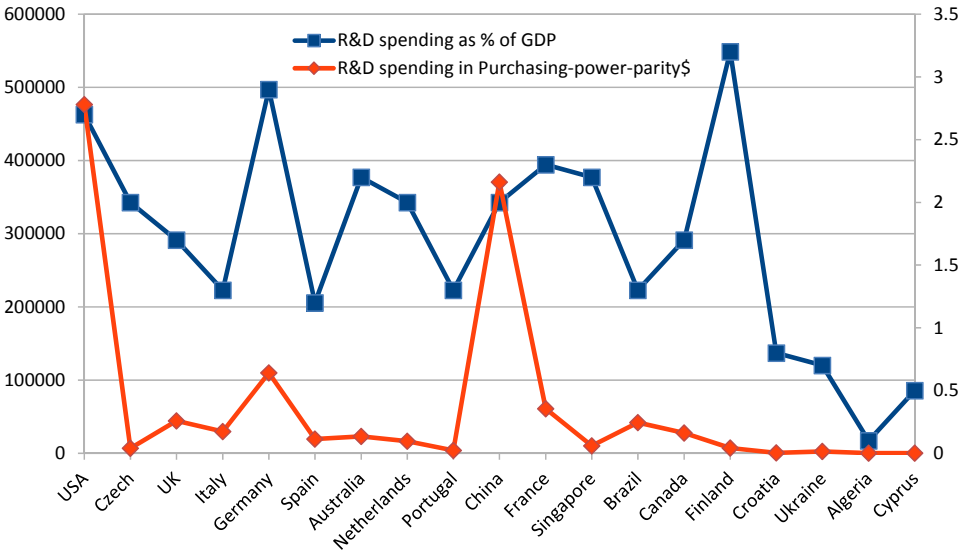


Figure A.6: The investments in R&D [290] of the publishing countries after the fine-grained exclusion/inclusion step

A.2/ Research Topics and Application Domains (SLRQ5)

This section discusses the results and answers the SLR questions raised in SLRQ5 (Section 4.2.1). In particular, it deals with the research topics (Section A.2.1) and the application domains (Section A.2.2).

A.2.1/ Research Topics (SLRQ5.1)

Subjects or issues that a researcher is interested in when conducting research on UAVs. These "research topics" provide the general directions to researchers for exploring, defining, and refining their ideas. Figure A.7 plots the main research topics addressed in the reviewed articles and Table A.1 lists the papers per each research topic. It is worth mentioning that some papers were flagged as having more than one research topic. Therefore, and to normalize the weights given by each paper to the research topics distribution, we have decided to assign the most dominant research topic per paper. The addressed research topics are:

1. Coordination (17%), UAVs interact to coordinate their actions for reaching their common objectives;

2. Mission Management (14%) addresses the optimal dynamic assignment of high-level missions, *i.e.* objectives, to UAVs. High-level missions are those where the UAVs rely on a high-level description of their objectives without many details and without human guidance;

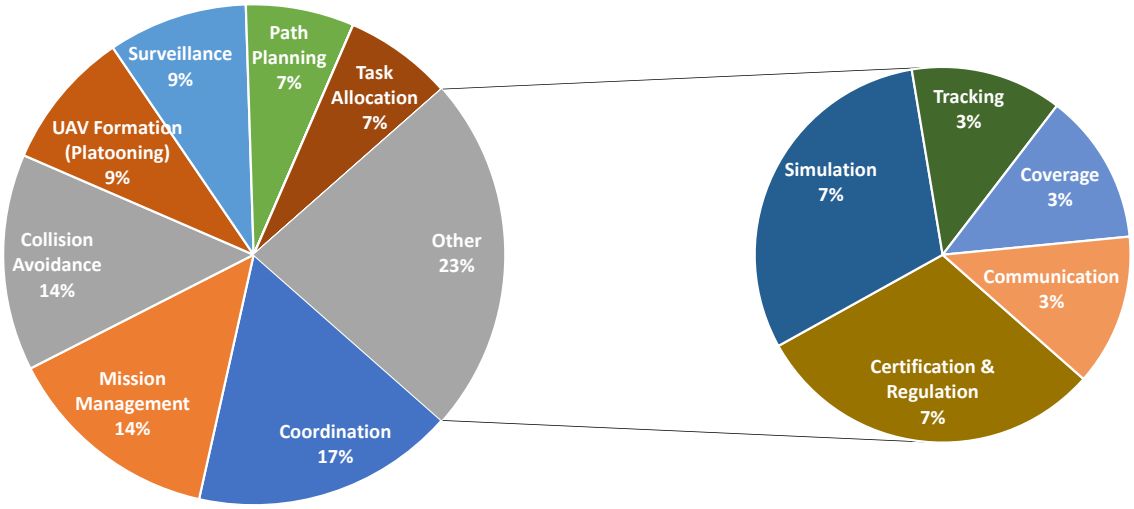


Figure A.7: The main research topics related to civilian UAVs applications

3. Collision Avoidance (14%) allows increasing UAV safety by avoiding collisions among UAVs, persons, animals, and other objects;
4. UAV Formation (Platooning) (9%) addresses the definition of flight formations for UAVs;
5. Surveillance (9%) enables UAVs to help the detection of dangerous and illegal situations;
6. Path Planning (7%) focuses on the static and dynamic computing of the best paths to fly along according to the environmental constraints;
7. Task Allocation (7%) is the optimal dynamic (potentially distributed) assignment of tasks to the UAVs;
8. Certification & Regulation (7%) is related to the definition of regulations dedicated to UAVs, and of the associated certifications for UAVs or pilots;
9. Simulation (7%) focuses on the design and implementation of developed simulation frameworks to understand and validate UAV behaviors;
10. Communication (3%) is related to the definitions of the means of communication between the UAVs, and between the UAVs and the ground infrastructure;
11. Coverage (3%) addresses the problems of map coverage;
12. Tracking (3%) focuses on the detection and tracking of objects in the environment of the UAVs.

Therefore, as Figure A.7 shows, the research topics have diverse aims and tackle several aspects related to UAVs, ranging from low-level (*i.e.* close to hardware) issues, such

Research Topic	Papers per Topic
Coordination	Bürkle et al. [49], Zhu et al. [325], Rollo et al. [243], De Benedetti et al. [73], Cimino et al. [65], Ciarletta et al. [64], Obdržálek [219]
Mission Management	Wei et al. [306], Gunetti et al. [114, 115], Stenger et al. [266], Sampedro et al. [248], Fulford et al. [98]
Collision Avoidance	Ashraf et al. [18], Kandil et al. [140], Šišlák et al. [263], Arokiasami et al. [17], Zou et al. [326], Kucherov and Kucherov [159]
UAV Formation (Platooning)	Van der Walle et al. [301], Sutton et al. [271], Benedetti et al. [31], Bürkle and Leuchter [48]
Surveillance	Semsch et al. [254], Khaleghi et al. [143], Bentz and Panagou [32], da Silva et al. [261]
Task Allocation	Wei et al. [307], Evertsz et al. [86], Vasilijevic et al. [292]
Path Planning	Volf et al. [299], Peng et al. [230], Ma et al. [181]
Certification & Regulation	Webster et al. [305, 304], Schatten [249]
Simulation	Pechoucek et al. [228], Veloso et al. [294], De Benedetti et al. [74]
Communication	Agogino et al. [7]
Coverage	Albani et al. [9]
Tracking	Ferrag et al. [90]

Table A.1: The reviewed papers per research topic

as UAV communication, to high-level concerns that require considerable UAV autonomy (*e.g.* mission management).

To assess the UAV autonomy involved in each research topic, we rely on an autonomy metric proposed by Clough [66] to measure Autonomous Control Levels (ACL) of UAVs. This metric is a scale ranging from 0 (for a remotely piloted non-autonomous UAVs) to 10 (for a fully autonomous UAVs). Figure A.8 plots the average ACL for each research topics. Note that ACL values were either mentioned explicitly by the authors of the primary studies or were determined by the reviewers by evaluating the UAV autonomy according to the ACL scale.

As seen in Figure A.8, some research topics tended to endow UAVs with more autonomy than others. For instance, coverage, coordination, surveillance, and mission management need more autonomy than path planning, collision avoidance, and communications. Note that research topics such as certification & regulation and simulation attained relatively high ACL. However, the main concerns of these works were building a simulation environment for the UAVs (in case of the simulation research topic) and certifying that UAVs adhere to the enforced regulations & norms (in case of the certification & regulation research topic). For this reason, the UAVs implementation provided by these works were mainly case-studies lacking details about the evaluations and the implementations.

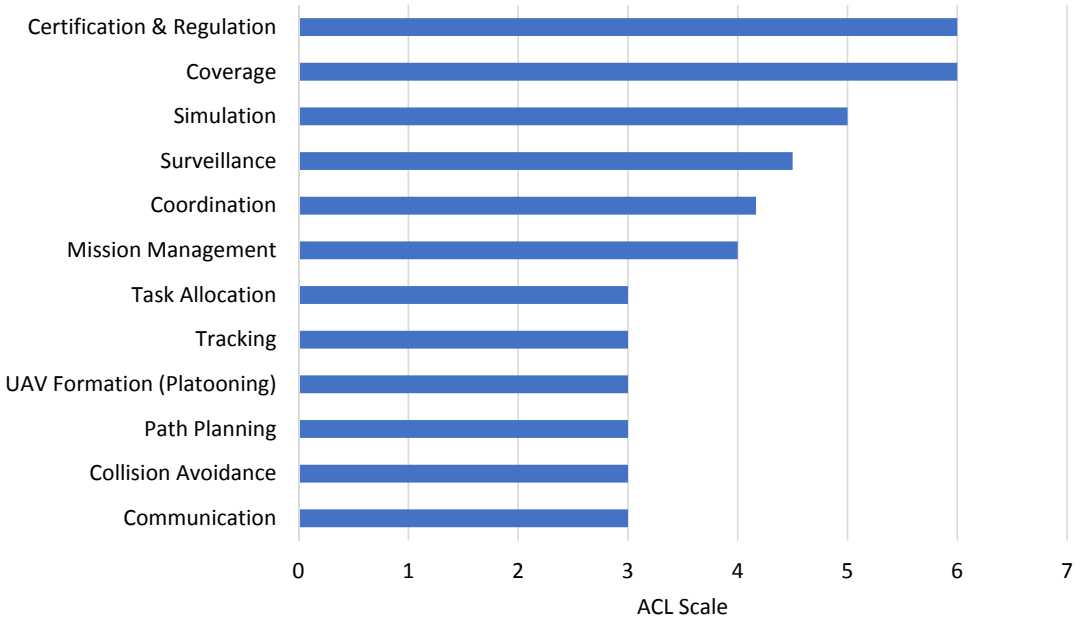


Figure A.8: The average ACL for each research topic

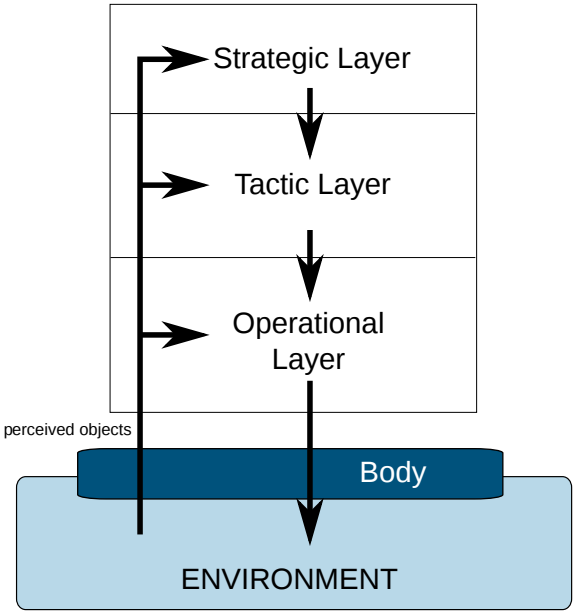
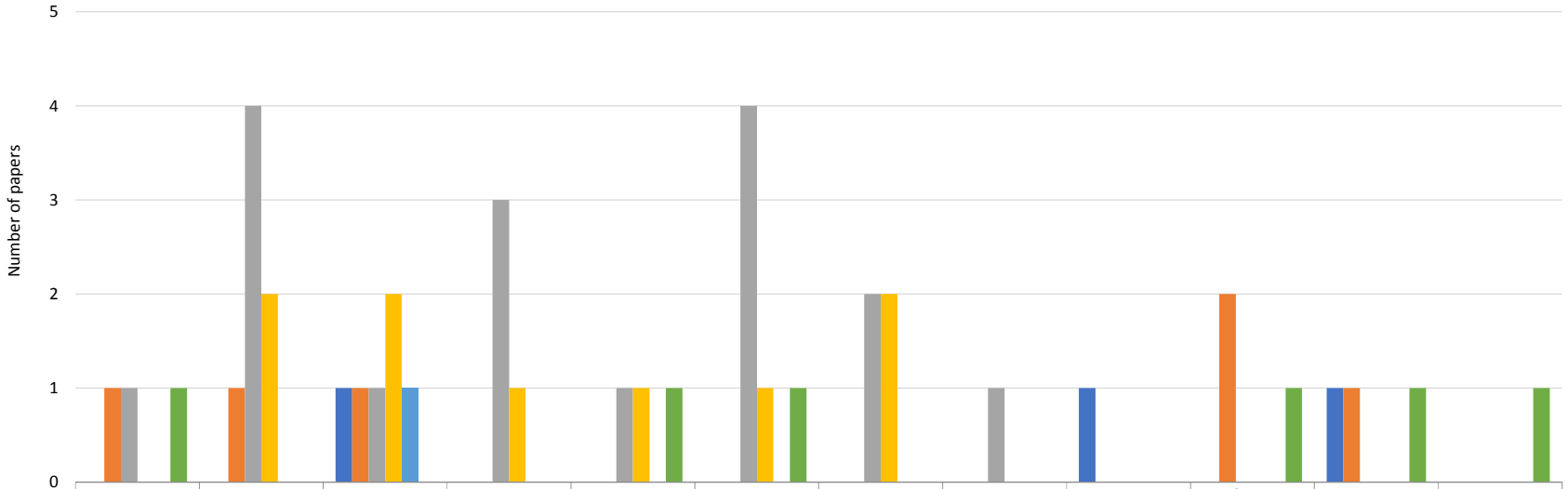


Figure A.9: The multi-layer architecture of the agents, adapted from [196]

Furthermore, no reviewer reported a paper with an ACL scale higher than 6. This is explained by the fact that higher levels of autonomy in the ACL scale were associated with specific military application requirements (*e.g.* battle-space knowledge, battle-space cognizance, *etc.*) and this SLR focuses on work related to civilian applications only. Note that this issue could be solved by relying on other metrics allowing to evaluate the maturity of the contributions. Technology Readiness Level (TRL) [197] is one metric that would help in this direction. Yet, unlike the ACL, it is not focused on UAV autonomy. For this reason, we opted for ACL in this paper.



	Task Allocation	Coordination	Collision Avoidance	Surveillance	Path Planning	Mission Management	UAV Formation (Platooning)	Coverage	Communication	Certification & Regulation	Simulation	Tracking
Evolutionnary	0	0	1	0	0	0	0	0	1	0	1	0
BDI	1	1	1	0	0	0	0	0	0	2	1	0
Cognitive	1	4	1	3	1	4	2	1	0	0	0	0
Reactive	0	2	2	1	1	1	2	0	0	0	0	0
Flocking	0	0	1	0	0	0	0	0	0	0	0	0
Unknown	1	0	0	0	1	1	0	0	0	1	1	1

Figure A.10: The number of papers with agent architecture as per the research topic

Figure A.10 shows the number of papers with agent architecture as per the research topic. The most used agent architectures are cognitive (without BDI): 17, and reactive (without flocking): 9, considering all the research topics. This fact is explained by the characteristics of the research topics: the ideal agent-based modeling of UAVs is usually based on a multi-layer architecture [196], illustrated in Figure A.9. The operational layer corresponds to the (very-)short term, *i.e.* the control of the UAVs. The tactic layer is associated with the planning of the UAV actions, *e.g.* path planning. The strategic layer is associated with the missions of the UAVs. In this layer, mission and task management need more complex models typically found in the cognitive scope. It is interesting to note that a low number of papers contains a multi-layer model. The other papers focus on a single layer, mostly tactic or strategic.

A.2.2/ Application Domains (SLRQ5.2)

In addition to research topics, SLRQ5 addresses application domains of UAVs. They refer to the applied research, in which scientific studies and research works aim to solve practical problems. Figure A.11 shows the distribution of the civilian UAV application domains of studied papers, and Table A.2 lists the papers per each application domain. It is worth mentioning that some papers tackle several application domains. Therefore, and to normalize the weights given by each paper to the application domains distribution, we have decided to assign the most dominant application domain per paper. For the paper that has no dominant application domain, or if its contribution is application-independent, it is considered to be in the General domain. The resulting application domains are:

1. General (53%);
2. Urban Planning (26%);
3. Emergency Response (12%);
4. Telecommunication (5%);
5. Agriculture (2%);
6. Border surveillance (2%).

As seen in Figure A.11, Urban planning and General are the most common application domains. This figure shows the growing attention given to UAV applications in civilian and urban environments since the share of urban planning is about 26%, while other application domains such as agriculture, emergency response, border surveillance, which often take place outside the urban environment, received less attention.

Application domain	Papers per application domain
General	Ashraf et al. [18], Wei et al. [307], Bürkle et al. [49], Wei et al. [306], Rollo et al. [243], Gunetti et al. [114, 115], Evertsz et al. [86], Webster et al. [304], Sampedro et al. [248], Sutton et al. [271], Peng et al. [230], Zou et al. [326], Ma et al. [181], Bentz and Panagou [32], da Silva et al. [261], De Benedetti et al. [73, 74], Obdržálek [219], Fulford et al. [98], Kucherov and Kucherov [159], Ferrag et al. [90]
Urban Planning	Semsch et al. [254], Volf et al. [299], Webster et al. [305], Kandil et al. [140], Arokiasami et al. [17], Pechoucek et al. [228], Veloso et al. [294], Šišlák et al. [263], Khaleghi et al. [143], Bürkle and Leuchter [48], Schatten [249]
Emergency Response	Zhu et al. [325], Van der Walle et al. [301], Benedetti et al. [31], Cimino et al. [65], Vasilijevic et al. [292]
Telecommunication	Agogino et al. [7], Ciarletta et al. [64]
Border Surveillance	Stenger et al. [266]
Agriculture	Albani et al. [9]

Table A.2: The reviewed papers per application domain

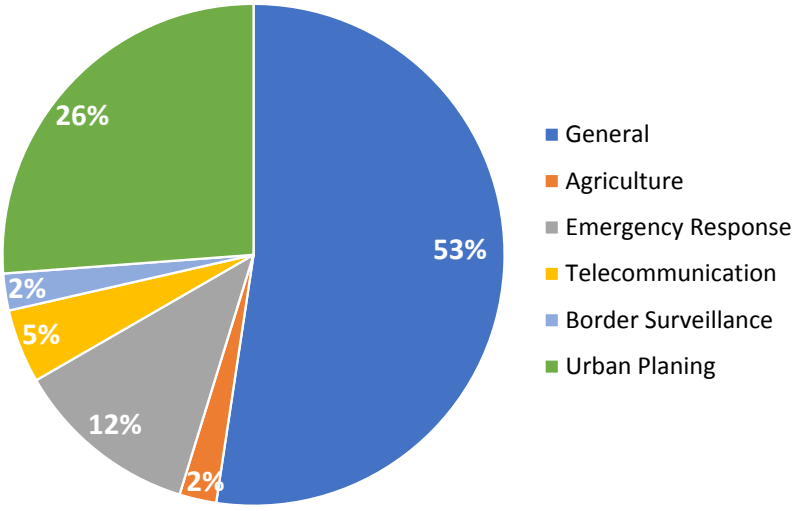


Figure A.11: The main civilian UAV application domains

To assess the maturity of the reviewed primary studies, we resorted to the quality criterion Q3 defined in Table 4.1. Q3 evaluates the quality of the experiments conducted by the authors of primary studies and the statements of the obtained results. The intuition here is that the more mature the application domain is the higher would be its score for Q3.

Figure A.12 shows the average Q3 score obtained per application domain. As can be seen from the figure, based on their maturity, the application domains can be classified into two clusters. The first cluster represents relatively mature application domains (agriculture, telecommunications, and emergency response). The second cluster represents less mature application domains (general, urban planning, and border surveillance). Note

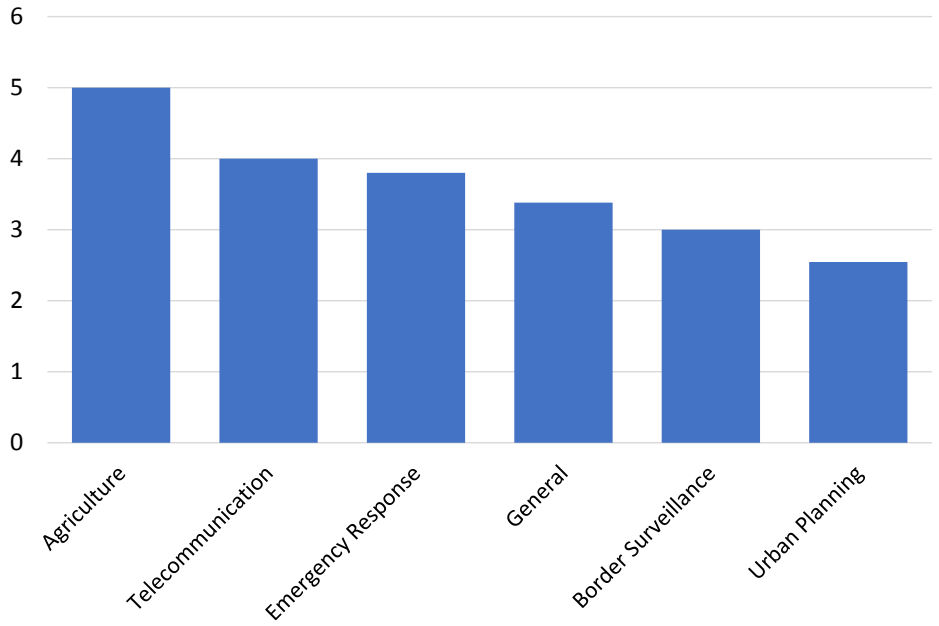


Figure A.12: The average of quality metric Q3 given by reviewers per the application domain

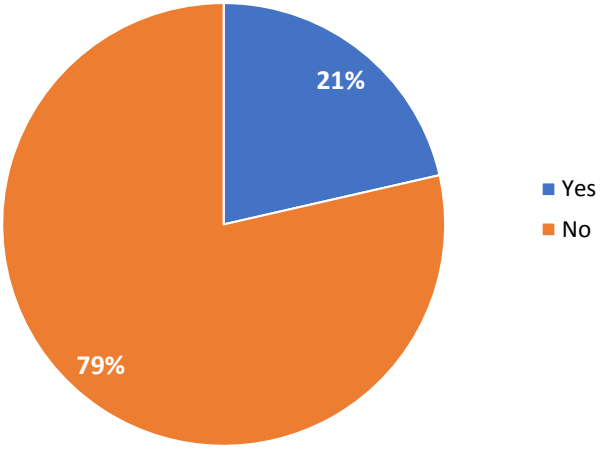


Figure A.13: The proportion of the models that are including or not the IoT concept

that these assessments only concern works using ABS for civilian UAV applications.

A.3/ Unmanned Aerial Vehicles with Internet Of Things (SLRQ6)

The smart city concept integrates Information and Communication Technology (ICT) (*cf.* Section SLRQ7), and various physical devices connected to the network, *e.g.* Internet Of Things (IoT) or Wireless Sensor Network (WSN) [106].

Figure A.13 shows the proportion of papers that include the IoT concept against the papers that do not. 9 papers include IoT, and 33 do not. According to [106], several opportunities for UAVs use to support a smart city exist. These opportunities will be very beneficial to any smart city that would utilize UAVs for their economic growth and development. Therefore, it is important to investigate whether IoT is considered in the reviewed models. The low proportion of reviewed papers that are considered IoT indicates that it is still an open issue. According to our knowledge, this proportion may be explained by the fact that researchers focus on the UAV behavior itself, not on the UAV environment.

A.4/ UNMANNED AERIAL VEHICLES COMMUNICATION (SLRQ7)

Vehicle-to-everything (V2X) communication is the passing of information from a vehicle to any entity that may affect the vehicle and vice versa [308]. It is a vehicular communication system that incorporates other more specific types of communication as Vehicle-to-Infrastructure (V2I), Vehicle-to-Network (V2N), Vehicle-to-Vehicle (V2V), Vehicle-to-Pedestrian (V2P), Vehicle-to-Device (V2D), and Vehicle-to-Grid (V2G). The purpose of this research question is to identify the V2X used by the UAVs to connect to other entities. Among all these types of communication, reviewers have found references to V2I and V2V only.

The main motivations for V2X are road safety, traffic efficiency, and energy savings within smart cities. UAVs are one of the means for setting up the smart city concept [106]. Therefore, it is important to highlight the V2X technologies that are considered within UAV models.

Figure A.14 plots the type of communications used: 33% of the papers includes com-

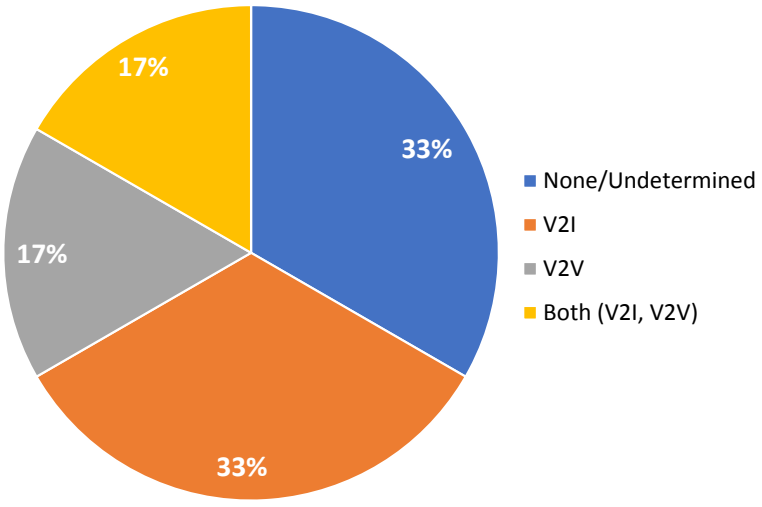


Figure A.14: The types of communications

munication between the UAVs and the infrastructure (V2I), 17% between UAVs (V2V) by using: *(i)* implicit communication, *i.e.* the communication means are not explicitly described; *(ii)* direct communication means, *e.g.* Wireless; or *(iii)* by using stigmergy communication. 17% of the papers propose a model with both V2I and V2V components.

Finally, 33% of the papers do not consider any specific communication approach. Therefore, when communication is considered within these primary studies, it is not detailed; and the authors seem to assume that the information is exchanged whatever the communication mean is. These papers are not excluded because they contain models that support interaction among the UAV agents, even if it is at an abstract level.

It is interesting to note that 66% of the papers consider that UAVs are connected entities that need to interact with their environment or with other UAVs. This is in-line with the fact that UAVs may contribute to set up the smart city concept (*cf.* Section SLRQ6).

A.5/ Evaluation and Simulation Scenarios (SLRQ8)

This research question is related to the evaluation of the proposed models (Figure A.15): whether it relies on a dataset, a generated synthetic dataset, or no dataset. Only a few of the reviewed papers (5%) have a reference to a dataset for setting up the UAV simulations; 17% of the papers have generated a specific dataset for evaluation; While, the majority have no dataset. In papers with no datasets, simulation scenarios are defined as ad hoc by authors. This relatively low number of papers with datasets may be explained by the difficulties of building such sets, *e.g.* it is hard to gather realistic data and initialize a UAV model from it. Having well-established testbeds with datasets helps to unify the testing process, and allows for systematic comparisons between the proposed solutions.

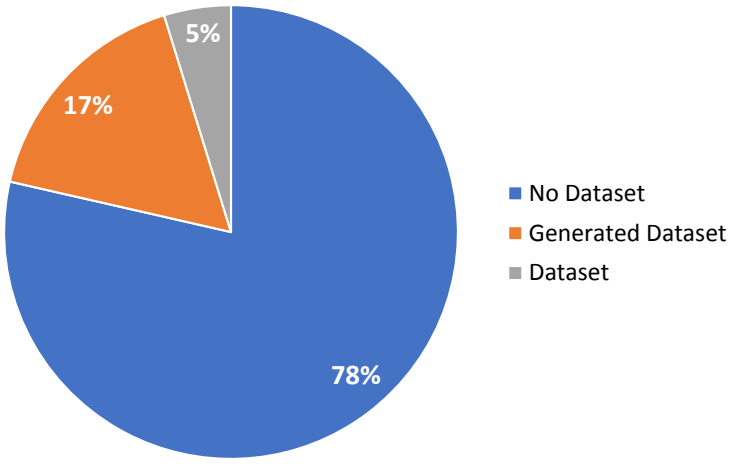


Figure A.15: The use of datasets by the reviewed papers

A.6/ CONCLUSION

In Chapter 4, and following a well-established SLR methodology, we identified 8 SLRQs helping to assess the contributions of MAS and ABS in civilian UAV applications. The main findings of the 5 SLRQs not included in Chapter 4 (SLRQ4, SLRQ5, SLRQ6, SLRQ7, SLRQ8) are[1]:

1. Research on MAS and ABS for civilian UAV applications has witnessed a considerable increase in the past decade and most of the reviewed papers were written in Europe followed by North America and Asia.
2. Coordination, mission management, UAV formation (platooning), collision avoidance, task allocation, and path planning were the most studied research topics while "Urban planning" and "General" accounted for the majority of application domains.
3. Despite the key role that the UAV is expected to assume in smart cities and connected smart environments, only a fifth of the reviewed paper integrate IoT technologies in their research works.
4. The majority of the reviewed papers address UAV connectivity. This shows that most of the reviewed papers view UAVs as connected entities both among themselves and with their environment.
5. To evaluate their contributions, only 5% of papers rely on public datasets and less than 20% use generated datasets. The remaining majority do not use any dataset for evaluation purposes. This underlines the absence of common testbeds and datasets allowing to evaluate and compare these works.

Regulation and collision avoidance are among the prominent challenges to be settled. Yet, the air is still largely unregulated and unmarked, especially to the naked eye, unequipped with height measuring methods, without prior knowledge of any restrictions regarding the filming of surrounding people, and the seriousness of the threat a UAV poses as it zooms past or above people. Moreover, enhanced availability of better Global Positioning System (GPS) trackers, quieter copters, and smaller "footprint" also raises new legal issues and requires current and up-to-date regulation. Nevertheless, the vast majority of the world still remains behind on effective UAV control. Yet, these issues have not received enough attention in the reviewed papers. Namely, regulation is addressed directly in no more than 7% of the papers. while collision avoidance and UAV safety are addressed only by 11% of the papers.

[1] *see* Section 4.3 for SLRQ1, SLRQ2, and SLRQ3.

B

Comparison of Agent-based Simulation Frameworks

B.1/ Introduction

In this appendix, and as a further detailed investigation of the Systematic Literature Review (SLR) question 3 (page 43), a comparison of frameworks for civilian UAV applications is provided. This appendix is based on the paper:

Yazan Mualla, Wenshuai Bai, Stéphane Galland, and Christophe Nicolle, *Comparison of Agent-Based Simulation Frameworks for Unmanned Aerial Transportation Applications*, Procedia Computer Science 130, Elsevier, pp. 791-796 (2018). DOI: 10.1016/j.procs.2018.04.137.

Recently, the civilian applications of Unmanned Aerial Vehicles (UAVs) in aerial transportation are gaining more interest. Due to operational costs, safety concerns, and legal regulations, Agent-based Simulation (ABS) frameworks are preferably used to implement models and conduct tests. This appendix introduces a methodology to compare the most widely used frameworks. The methodology is inspired by the International Organization for Standardization (ISO) software quality model and uses a weighted sum scoring system to give points to the frameworks under study. The proposed criteria in the methodology consider ABS features and adapt specific features of unmanned aerial transportation. Preliminary comparison results and recommendations are provided and discussed. This comparison helped us in choosing the framework to adopt for the pilot test in Chapter 8.

Despite the promising research efforts of ABS in the domain of UAVs, very few works were dedicated to understand and analyze existing works using ABS in civilian UAV applications. Very few surveys outlined a comprehensive set of research questions pertaining to multi-agent simulations for civilian UAV applications. There are works comparing frameworks in the literature. Nonetheless, these works either:

1. address other applications such as energy consumption, geo-spatial applications, or parallel & distributed applications, or

2. focus on measuring and assessing the performance of frameworks.

Against this background, the objective of this appendix is to rely on the results of the SLR to in Chapter 4 analyze existing frameworks by selecting specific criteria to evaluate them based on civilian UAV application considerations.

The rest of this appendix is organized as follows: Section B.2 discusses the other comparisons in the literature. The main work is provided in Section B.3. First, the software quality model from the literature of software comparison is discussed in Section B.3.1. Second, the general features of the frameworks under study are provided in Section B.3.2. Third, Section B.3.3 details the ranking criteria we defined to compare the frameworks, and explains the weighted sum scoring system. Fourth, Section B.3.4 lists and discusses the results of the comparison. Section B.4 concludes this appendix.

B.2/ Other Comparisons in the Literature

Several surveys about the comparison of ABS frameworks were proposed. Railsback et al. [236] implemented a simple scenario with 100 agents randomly moving on a small grid (100x100 cells) for measuring the performance of frameworks. Five platforms (NetLogo [313], MASON [180], Repast Simphony [68], Java Swarm and Objective-C Swarm) were compared. The authors identified future priorities encouraging researchers and developers to adopt an agent-based framework and improve their performance. Their results were mostly limited to their perception and experience.

Abar et al. [1] presented a comprehensive comparative survey of 85 frameworks. However, the comparison criteria included only general features and some frameworks were not included like Gazebo [153] and AirSim [256] which we find related to UAV applications (*cf.* Figure 4.10). Lorig et al. [179] provided a technical comparison of 4 frameworks, but they focused only on one metric which is scalability. In [285], the aim was to determine the framework best suited for theory and data-based modeling of social interventions out of 4 candidates only. the authors have categorized various characteristics of the ABS toolkits into user-friendly taxonomies. The evaluation was based on official program documentation, statements by developers and users, and the experiences and impressions of the evaluators. The evaluation results showed the Repast Simphony [68] environment to be the clear winner.

Another survey by Serenko et al. [255] tackled the concepts of agent designing, modeling, and simulation toolkits available in the market. Their data collection efforts comprised the

download and trial use of 20 agent toolkits. The paper concluded that no single uniform toolkit satisfies the needs of all agent-related courses. Other surveys were reported by Kravari and Bassiliades [156], Nikolai and Madey [216], and Gupta and Kansal [118]. Although these surveys presented insightful information about ABS, they were mainly far from complete, and none of them focused on UAVs. In our paper, a wider comparison in terms of criteria is defined, with special attention paid to features related to UAV applications like the physics and the environment.

Some works compared the frameworks based on their results in a specific application domain not related to UAVs such as energy consumption applications [324, 157], geospatial applications [53], computational science applications [11], marketing applications [215], or parallel and distributed applications [283, 245]. These works were restricted by the constraints of the specific application domain they consider, which are not necessarily related to the UAV applications constraints. In terms of comparing frameworks based on UAV applications, there are two surveys: First, Vogeltanz [297] provided a survey of more than 50 free software for the design, analysis, modeling, and simulation of UAVs. However, the survey did not include all the free frameworks, and the selection of the frameworks has been focused on small subsonic UAVs. Second, Craighead et al. [71] presented a survey of computer-based simulators for unmanned vehicles including UAVs, but their work is outdated. Nonetheless, the survey concluded that it is no longer necessary to build a new simulator from scratch. Both these surveys did not consider the concepts of agents in their comparison criteria.

B.3/ Comparison of Frameworks

B.3.1/ Software Quality Model

To compare two frameworks, there is a need to measure the quality model of each framework. In the literature of software comparison, there is no consensus on one methodology for assessing software quality. Several efforts have been conducted by researchers to define a model for software comparison. Behkamal et al. [28] gathered most of the widely used models as follows:

1. McCall Model [93]: The main idea is to establish a relationship between quality features and metrics. Nonetheless, metrics are not necessarily objective in this model.

2. FURPS Model [144]: It organizes features in two different categories of requirements: Functional requirements defined by the input and the predicted output, and Non-functional requirements which are usability, reliability, performance, and sup-

portability. However, this model does not take into consideration some key features like the portability of software products.

3. Dromey Model [81]: It seeks to enhance the relationship between the features and the sub-features of software quality. One key disadvantage of this model that it lacks the criteria for measurement of software quality.

4. Bayesian Belief Network (BBN) Model [265]: it is organized as a hierarchical tree-like structure. The root of the tree is the 'Quality' node connected to quality features nodes, and each node is connected to corresponding quality sub-features. It allows designing complex models that are difficult to design with other models.

5. ISO Model [135]: The software product quality criteria are classified in a hierarchical tree structure of 6 features. These features are further classified into 21 sub-features. This model provides the widest range of comparison criteria in various aspects.

The ISO model is adopted in this work as it includes several interesting criteria in the context of the proposed comparison like for example: "Interoperability", "Compliance", "Understandability", "Learnability", "Operability", "Adaptability", "Installability", *etc*. However, some criteria are excluded as "Efficiency" as run-time comparison tests have not been performed in this work. Moreover, and derived from the SLR review, we introduce some extra criteria that are specifically important for UAVs like the representation of gravity and magnetic fields, environment dynamics, and the support of the force-based motion. The focus is on the criteria that are mostly related to and associated with flying objects in the space and their environment.

B.3.2/ Frameworks General Features

The frameworks to be compared are as follows: Gazebo [153], AirSim [256], Janus [101, 100], Repast Simphony [68], NetLogo [313], Flame [149], JADE [30], and MASON [180]. Table B.1 provides their general features. Some other frameworks that resulted from the SLR (Chapter 4) like AgentFly [264] and A-globe [262] are not considered in this comparison either because the framework is not open source or because the main applications that the framework supports are not civilian. Additionally, Simulink [223] is not considered, even though it is used by a considerable portion of reviewed papers (*cf.* Figure 4.10) because it is not purely an ABS framework. On the other hand, the results shown in Figure 4.10 shows that AirSim [256] has not been used in any of the papers reviewed in the SLR (Chapter 4) mainly because it is a new framework (2017). However, it is added to the comparison because it holds some potentials for UAV simulation, as it is a specialized framework for unmanned vehicles including UAVs.

	Gazebo [153]	AirSim [256]	Janus [101, 100]	Repast Simphony [68]	NetLogo [313]	Flame [149]	JADE [30]	MASON [180]
Main domain	Robots	Autonomous vehicles	General	General	General	General	General	General
License	Apache 2.0	MIT	Apache 2.0 (Janus), proprietary license (Jasim)	New BSD	GPL	GNU, Academic license	LGPL v2	Academic License
Open source	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
First release	2002	2017	2008	2003	1999	2006	2003	2005
Development status	Active	Active	Active	Active	Active	Active	Active	Active
Programming languages	C++	C++, Python, C#, Java	SARL, Java, JavaScript	Java, C#, C++, Visual Basic.Net, Lisp, Prolog, and Python	LOGO	C	Java	Java
Operating systems	Linux	Windows, Linux	Any with JVM	Any with JVM	Any with JVM	Most with C support	Any with JVM	Any with JVM

Table B.1: Frameworks General Features

B.3.3/ RANKING CRITERIA

The evaluation of the frameworks is divided into 4 categories, with each category having different criteria. Simple ranking by giving ranks to the frameworks is impractical, as there is a need to quantify the level each framework is achieving a criterion. Therefore, a comparison system based on scoring points is used. S_j represents the set of all points given to the options that a criterion j can take, where $S_j \subset \mathbb{N}$ and $\left|S_j\right| \geq 2$. For example, if a criterion j has three options: Option1 (0p), Option2 (1p) and Option3 (2p), then $S_j = \{0, 1, 2\}$. The criteria per category with the distribution of points are listed in the following:

A. *System Features:* The main system features of the framework.

1. Agent architecture: Support of either reactive or proactive agent architectures (1p); Hybrid support of both reactive and proactive architectures (2p).
2. Support the communication between agents: No communication (0p); Communication through the environment (indirect) or direct communication (1p); Indirect and direct communication between agents (2p).
3. User support (documentation, mailing list, defect list, tutorials, forum, examples, FAQs, wiki, API, *etc.*): Average (0p); Good (1p); High (2p).
4. Support of sensors that are embedded in the UAV: No (0p); Yes (1p).
5. Support of GIS: No (0p); Yes (1p).

B. *Operation and Execution:* The main operational and execution features of the framework.

1. Installation ease: The installation method and the required Information Technology skills to install. The simpler the method for the user is considered the better as follows: Command-line install (0p); Graphical User Interface (GUI) installer (1p).
2. Operational ease: The support of features like Integrated Development Environment, command prompt, click & point, syntax coloring, auto-completion, creation wizards, drag-and-drop, automatic code generation, *etc*. The evaluation is as follows: Inadequate (0p); Average (1p); Good (2p); High (3p).
3. Interaction with objects in the air, *i.e.* Vehicle-to-Vehicle (V2V) and the ground, *i.e.* Vehicle-to-Infrastructure (V2I): No interaction (0p); V2V or V2I (1p); V2V and V2I (2p).

C. *Environment:* The features of the reproduced environment and how the UAV interacts with it.

1. Environmental model: No model (0p); One type of models, either 2d or 3d (1p); Two types of models (2p).

2. Support of endogenous dynamic related to the environment objects: The ability to reproduce the objects that can be found in a city environment like traffic lights or objects connected to the Internet. The evaluation is as per the support of those objects as follows: No support (0p); Simple dynamic environment objects with the algorithm being part of the simulator (1p); Complex dynamic environment objects with the support of co-simulation (2p).
3. Simulation of ecological/environmental dynamics: The ability to simulate environmental dynamics like wind and rain of different intensities, different visibility conditions, *etc*. The evaluation is as follows: No support (0p); Simple simulation of environmental dynamics (1p); Complex simulation of environmental dynamics, possibly with a co-simulator (2p).

D. *Physics:* The simulation of the laws of physics in a realistic manner, and how well the exact position of an object in the space is determined.

1. Representation of gravity field: No (0p); Yes (1p).
2. Representation of magnetic field: No (0p); Yes (1p).
3. Collision avoidance algorithm: No (0p); Yes (1p).
4. Support of force-based motion: No (0p); Yes (1p).

Since the importance of a criterion depends on the application domain, we associate a weight to each criterion to reflect its importance for a given domain. The benefit of using weights is two-fold:

1. The weights assigned by the experts in the application domain as they are mainly application-dependent; or
2. Weights can be used to normalize the importance of the proposed criteria (*i.e.* ensure that all the criteria have the same impact on the total score), as the maximum score of one criterion may be different from the maximum score of another criterion. This is needed when there is no specific application domain (*i.e.* the comparison is done in general).

Equation B.1 defines the normalized weight for each criterion.

$$w_j = \frac{e_j}{\max(S_j)} \text{ such that } \sum_{a \in [1;C]} e_a = C \tag{B.1}$$

Where: w_j is the weight of criterion j; e_j is a coefficient provided by a domain expert for criterion j, and it represents the importance of the criterion for the specific domain; $\max(S_j)$ is the maximum score of criterion j; C is the total number of criteria. The domain

expert has a total of C points to be divided according to the importance of the criteria. In case the comparison is performed in general (*i.e.* without a specific application domain), there is no need for the domain expert to provide e_j values, *i.e.* $e_j = 1$ for all criteria; so the weights are only used in this case to normalize the scores of criteria.

For each framework, the total weighted score is calculated from the following equation.

$$T_i = \sum_{j=1}^{C} w_j . s_{ij} \tag{B.2}$$

Where: T_i is the total weighted score of the framework i; s_{ij} is the score of the framework i for criterion j.

B.3.4/ Results and Discussion

Table B.2 lists the results of the comparison of all the frameworks. This table is filled based on our knowledge and the documents and tutorials provided by their developers. The last column represents the maximum score a framework can achieve for a specific criterion. For each category, the weighted scores of all frameworks are calculated, and the total weighted scores are provided at the end (the last row). As can be seen from the table that Repast Simphony [68] achieves the highest total weighted score with 13.00/15.00 to be the best framework for the ABS of civilian UAV applications in general, with the best runner-up to be Gazebo [153]. However, it is not the best framework in all categories if considered individually. For instance, and considering only the Environment category, the best framework is Gazebo [153].

It is worth mentioning that the frameworks' scores are with tight differences, hence a small change in the values of the table, for example when a framework is updated in a future version, the order of the frameworks may change. This means that generally, the frameworks have somehow similar characteristics in terms of UAV simulation and that when one excels in a specific domain the other excels in another. The weights are used to normalize the importance of the criteria.

B.4/ Conclusion

In this appendix, a methodology to compare ABS frameworks has been defined focusing on features of civilian UAV applications. The preliminary results show that Repast Simphony [68] is the most suitable framework for simulating UAVs civilian applications. The runner-up, with a slightly lower score, is Gazebo [153]. From this result, the choice was made to adopt Repast Simphony [68] for the pilot test in the thesis (Chapter 8).

		Repast Simphony	Gazebo	NetLogo	AirSim	MASON	Janus	Flame	JADE	w_j	Max Score
A	Agent architecture	2	2	2	2	2	2	2	2	1/2	2
A	Communication between agents	1	1	1	1	1	2	2	2	1/2	2
A	User support	2	1	2	0	1	1	1	2	1/2	2
A	Sensors	1	1	1	1	1	1	1	1	1/1	1
A	GIS	1	1	1	1	1	1	1	0	1/1	1
	System Features Weighted Score	**4.50**	**4.00**	**4.50**	**3.50**	**4.00**	**4.50**	**4.50**	**4.00**		**5.00**
B	Installation ease	1	0	1	0	1	0	0	1	1/1	1
B	Operational ease	3	2	3	2	2	2	3	2	1/3	3
B	V2V/V2I	2	2	2	2	2	2	2	2	1/2	2
	Operation & Execution Weighted Score	**3.00**	**1.67**	**3.00**	**1.67**	**2.67**	**1.67**	**2.00**	**2.67**		**3.00**
C	Environmental model	2	2	2	1	2	2	2	1	1/2	2
C	Dynamic environment objects	2	2	1	2	1	2	1	2	1/2	2
C	Ecological/ Environmental dynamics	1	2	1	2	1	1	1	0	1/2	2
	Environment Weighted Score	**2.50**	**3.00**	**2.00**	**2.50**	**2.00**	**2.50**	**2.00**	**1.50**		**3.00**
D	Gravity	1	1	1	1	1	1	1	1	1/1	1
D	Magnetic	0	1	1	1	0	0	0	1	1/1	1
D	Collision avoidance	1	1	1	1	1	1	1	1	1/1	1
D	Force-based motion	1	1	0	1	1	1	1	0	1/1	1
	Physics Weighted Score	**3.00**	**4.00**	**3.00**	**4.00**	**3.00**	**3.00**	**3.00**	**3.00**		**4.00**
	TOTAL WEIGHTED SCORE	**13.00**	**12.67**	**12.50**	**11.67**	**11.67**	**11.67**	**11.50**	**11.17**		**15.00**

Table B.2: Frameworks General Comparison

C

Questionnaire of the Pilot Test

Explainable Artificial Intelligence (XAI): Package delivery using drones (12 questions)

This questionnaire follows the sequences of scenarios shown in the simulation. In this simulation, some drones work on behalf of you (the owner) to deliver packages of clients. This questionnaire is built to collect opinions on the simulation tool. All data will be collected and used adhering to the EU General Regulation on Data Protection.

* Required

Q0: Please choose your group*

Mark only one oval.

○ Group A
○ Group B
○ Group C

C.1/ Participant Details

C.1.1/ General

Q1: Gender

Mark only one oval.

○ Female
○ Male

Q2: Age

_____

Q3: Level of English language*

Mark only one oval.
○ A1
○ A2
○ B1
○ B2
○ C1
○ C2

Q4: What is your prior knowledge about drones?*

Mark only one oval.
○ 5 (Very good)
○ 4 (Good)
○ 3 (Neutral)
○ 2 (Low)
○ 1 (Very low)

Q5: Professional status*

Mark only one oval.
○ Undergraduate Student (Skip to question **Q5-1-1**)
○ Graduate (Skip to question **Q5-2-1**)

C.1.2/ UNDERGRADUATE

Q5-1-1: Year :

Mark only one oval.
○ Year 1: Common Core (TC01/TC02)
○ Year 2: Common Core (TC03/TC04)
○ Year 3: Branch (BR01/BR02)
○ Year 4: Branch (ST40/BR04)
○ Year 5: Branch (BR05/ST50)

Skip to question **Q6**

C.1.3/ GRADUATE

Q5-2-1: Years after graduating :

Mark only one oval.

○ 1
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7
○ 8
○ 9
○ 10+

Q5-2-2: Work Title :

Mark only one oval.

○ PhD student
○ Doctor
○ Other: ————————-

C.2/ FUNCTIONALITIES

Q6: Approximately how many packages were delivered in all the scenarios?*

————————-

Q7: Approximately how many problems (unexpected events) happened in all the scenarios?*

————————-

Q8: What is the maximum number of drones that you think you can follow as an operator?*

————————-

C.3/ STATISTICAL ANALYSIS

Q9: Do you believe the only one time you watched the simulation tool working was enough to understand it?*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Q10: How do you rate your understanding of how the simulation tool works?*

Mark only one oval.
○ 5 (Very high)
○ 4 (High)
○ 3 (Normal)
○ 2 (Low)
○ 1 (Very low)

Q11: Why do you think and explanation of the simulation tool is important?

Check all that apply.
□ I want to know what the AI just did.
□ I want to know that I understand this AI system correctly.
□ I want to understand what the AI will do next.
□ I want to know why the AI did not make some other decision.
□ I want to know what the AI would have done if something had been different.
□ I was surprised by the AI's actions and want to know what I missed.

Q12: The explanation of how the simulation tool works in the last sequence has too many details.*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Thanks for submitting a response. We appreciate your help.

D

Questionnaire of the Main Test

Explainable Artificial Intelligence (XAI): Package delivery using drones (21 questions)

This questionnaire follows the sequences of scenarios shown in the simulation. In this simulation, some drones work on behalf of you (the owner) to deliver packages of clients. This questionnaire is built to collect opinions on the simulation tool. All data will be collected and used adhering to the EU General Regulation on Data Protection.

* Required

Q0: Please choose your group*

Mark only one oval.

○ Group A
○ Group B
○ Group C

D.1/ Participant Details

D.1.1/ General

Q1: Gender

Mark only one oval.

○ Female
○ Male

Q2: Age

_____

Q3: Level of English language*

Mark only one oval.
○ A1
○ A2
○ B1
○ B2
○ C1
○ C2

Q4: What is your prior knowledge about drones?*

Mark only one oval.
○ 5 (Very good)
○ 4 (Good)
○ 3 (Neutral)
○ 2 (Low)
○ 1 (Very low)

Q5: Professional status*

Mark only one oval.
○ Undergraduate Student (Skip to question **Q5-1-1**)
○ Graduate (Skip to question **Q5-2-1**)

D.1.2/ UNDERGRADUATE

Q5-1-1: Year :

Mark only one oval.
○ Year 1: Common Core (TC01/TC02)
○ Year 2: Common Core (TC03/TC04)
○ Year 3: Branch (BR01/BR02)
○ Year 4: Branch (ST40/BR04)
○ Year 5: Branch (BR05/ST50)

Skip to question **Q6**

D.1.3/ GRADUATE

Q5-2-1: Years after graduating :

Mark only one oval.

○ 1
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7
○ 8
○ 9
○ 10+

Q5-2-2: Work Title :

Mark only one oval.

○ PhD student
○ Doctor
○ Other: ————————-

D.2/ FUNCTIONALITIES

Q6: Approximately how many packages were delivered in all the sequences?*

————————-

Q7: Approximately how many problems (unexpected events) happened in all the sequences?*

————————-

Q8: What is the maximum number of drones that you think you can follow as an operator?*

————————-

D.3/ STATISTICAL ANALYSIS

Q9: The number of drones (10 drones) in the last scenario was not overwhelming (too much to follow).*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Q10: Do you believe the only one time you watched the simulation tool working was enough to understand it?*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Q11: How well the simulation tool helped you to understand how it works?*

Mark only one oval.
○ 5 (Very much)
○ 4 (Good enough)
○ 3 (I'm neutral about it)
○ 2 (Not good enough)
○ 1 (Not at all)

Q12: How do you rate your understanding of how the simulation tool works?*

Mark only one oval.
○ 5 (Very high)
○ 4 (High)
○ 3 (Normal)
○ 2 (Low)
○ 1 (Very low)

Q13: I am confident in the simulation tool. I feel that it works well.*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Q14: The outputs of the simulation tool are very predictable.*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Q15: The simulation tool is very reliable. I can count on it to be correct all the time.*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Q16: The simulation tool is efficient in that it works very quickly.*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Q17: I am wary of the simulation tool.*

Mark only one oval.
○ 5 (I must stay and keep an eye on it)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)

○ 1 (I can leave it to work by itself)

Q18: Why do you think and explanation of the simulation tool is important?

Check all that apply.
□ I want to know what the AI just did.
□ I want to know that I understand this AI system correctly.
□ I want to understand what the AI will do next.
□ I want to know why the AI did not make some other decision.
□ I want to know what the AI would have done if something had been different.
□ I was surprised by the AI's actions and want to know what I missed.

Q19: From the explanation, I understand better how the simulation tool works.*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Q20: The explanation of how the simulation tool works is satisfying.*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Q21: The explanation of how the simulation tool works in the last sequence has sufficient details.*

Mark only one oval.
○ 5 (I agree strongly)
○ 4 (I agree somewhat)
○ 3 (I'm neutral about it)
○ 2 (I disagree somewhat)
○ 1 (I disagree strongly)

Thanks for submitting a response. We appreciate your help.

E

Box Plots of Insignificant Results in the Main Test

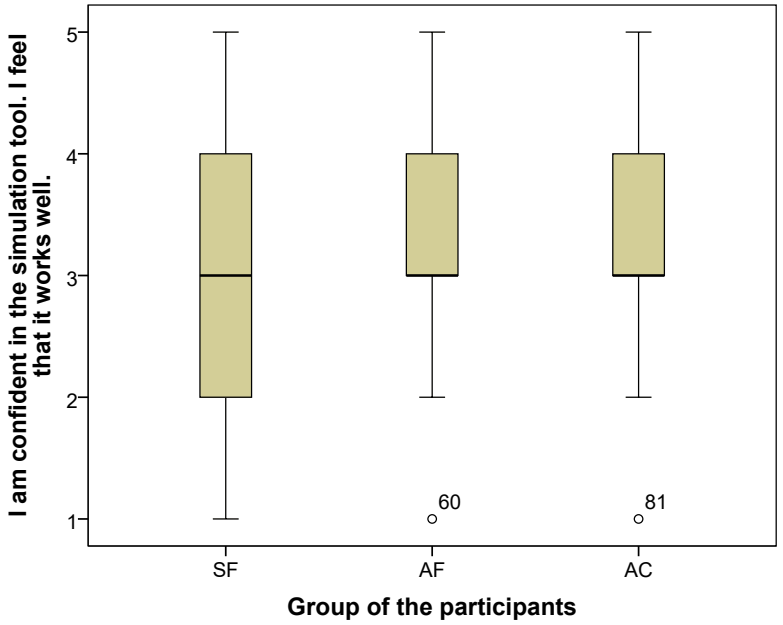


Figure E.1: Main test: Q13 ($\overline{x}_{SF} = 2.83, \overline{x}_{AF} = 3.20, \overline{x}_{AC} = 3.40$, medians are represented in the figure)

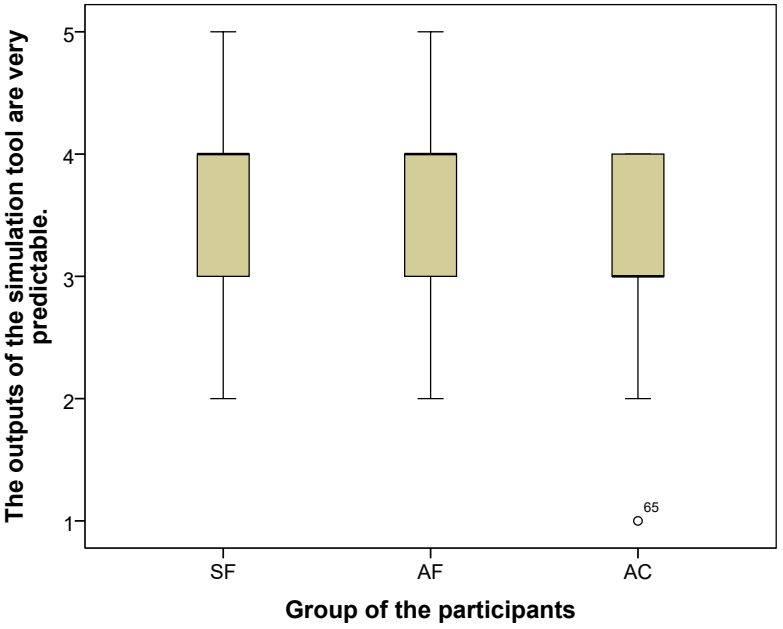


Figure E.2: Main test: Q14 ($\overline{x}_{SF} = 3.63, \overline{x}_{AF} = 3.67, \overline{x}_{AC} = 3.20$, medians are represented in the figure)

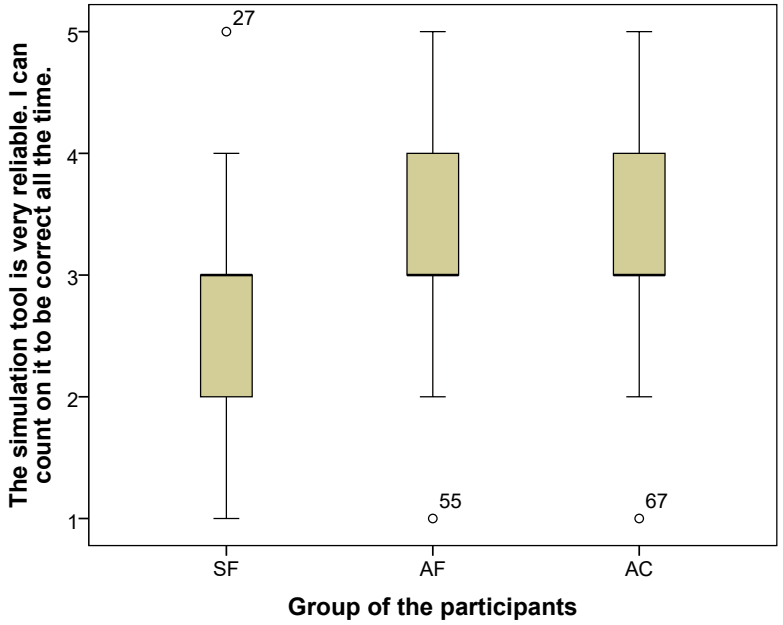


Figure E.3: Main test: Q15 ($\overline{x}_{SF} = 2.83, \overline{x}_{AF} = 3.13, \overline{x}_{AC} = 3.30$, medians are represented in the figure)

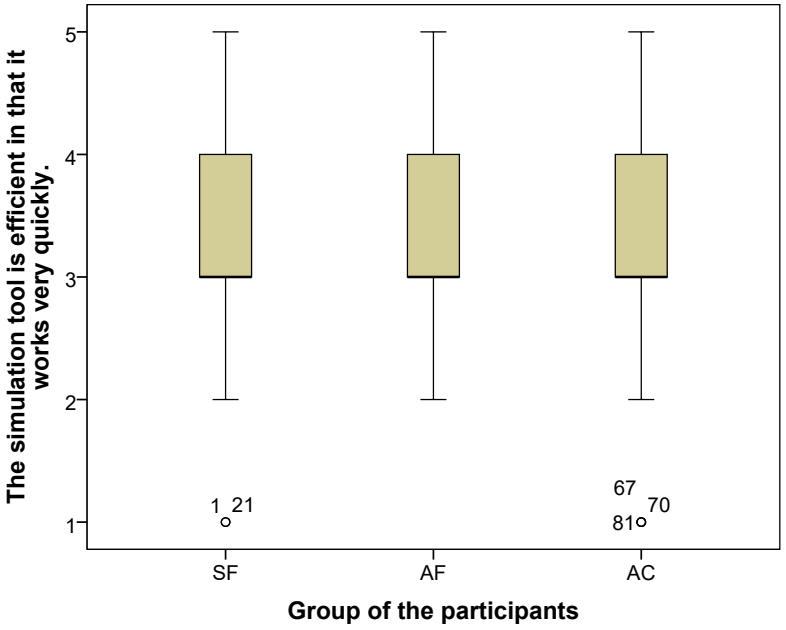


Figure E.4: Main test: Q16 ($\overline{x}_{SF} = 3.20, \overline{x}_{AF} = 3.33, \overline{x}_{AC} = 3.17$, medians are represented in the figure)

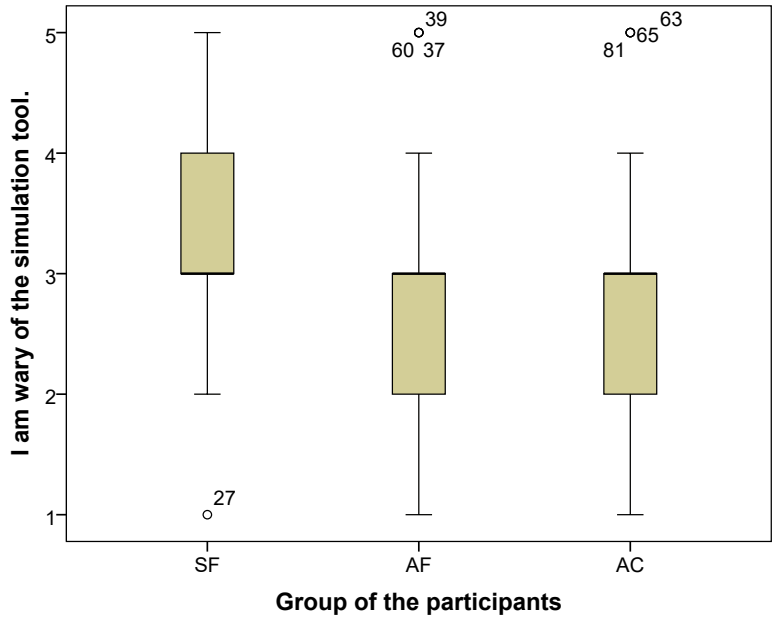


Figure E.5: Main test: Q17 ($\overline{x}_{SF} = 3.40, \overline{x}_{AF} = 2.97, \overline{x}_{AC} = 2.90$, medians are represented in the figure)

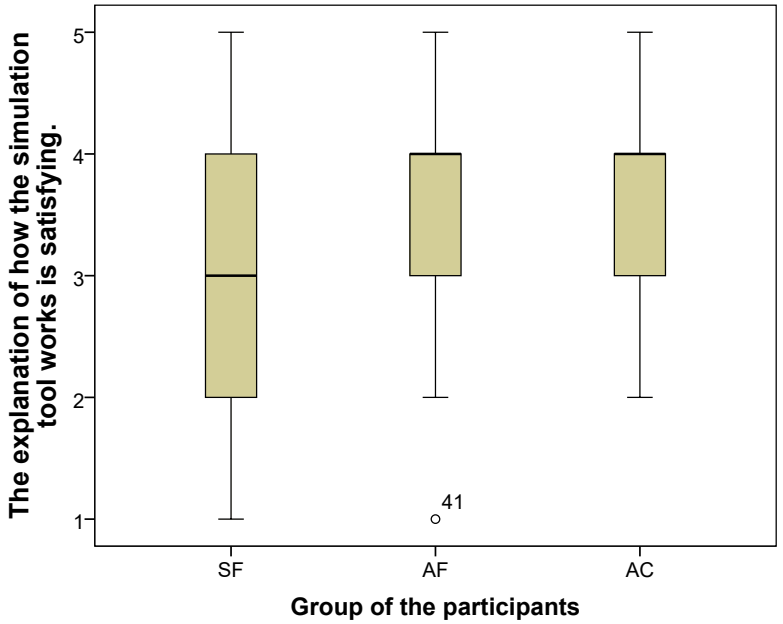


Figure E.6: Main test: Q20 ($\overline{x}_{SF} = 3.10, \overline{x}_{AF} = 3.53, \overline{x}_{AC} = 3.73$, medians are represented in the figure)

F

PUBLICATIONS OF THE AUTHOR

In this appendix, the list of the publications of YAZAN MUALLA during the Ph.D. is provided. The following table shows the number of publications per type:

Number of publications by YAZAN MUALLA	
International journal with reading committee	3
International conference with proceedings	5
National conference with proceedings	4
International workshop with proceedings	7
International conference/workshop without proceedings	4
Total	**23**

The publications are listed in two categories: (i) publications that are directly related to the content of this Ph.D. thesis, and (ii) publications that are not directly related to the content of the thesis, but are in the domain of interest of the thesis, and to which YAZAN MUALLA has provided a concrete contribution.

F.1/ PUBLICATIONS DIRECTLY RELATED TO THE PH.D. THESIS

Number of publications related to this Ph.D. thesis	
International journal with reading committee	2
International conference with proceedings	2
International workshop with proceedings	3
International conference/workshop without proceedings	4
Total	**11**

INTERNATIONAL JOURNAL WITH READING COMMITTEE

- Yazan Mualla, Amro Najjar, Alaa Daoud, Stéphane Galland, Christophe Nicolle, Ansar-Ul-Haque Yasar, and Elhadi Shakshuki,

Agent-Based Simulation of Unmanned Aerial Vehicles in Civilian Applications: A Systematic Literature Review and Research Directions,
International Journal of Future Generation Computer Systems (SJR Quartile: **Q1**, Clarivate Analytics Impact Factor 2019: **6.125**), Elsevier, vol. 100, pp. 344-364 (2019). DOI: 10.1016/j.future.2019.04.051

- Stéphane Galland, Yazan Mualla, Igor Tchappi Haman, Hui Zhao, Sebastian Rodriguez, Amro Najjar, and Nicolas Gaud,
Model Transformations from the SARL Agent-Oriented Programming Language to an Object-Oriented Programming Language,
International Journal on Agent-Oriented Software Engineering, vol. 7, pp. 37-75 (2019). DOI: 10.1504/IJAOSE.2019.106458

International Conference with Proceedings

- Yazan Mualla, Amro Najjar, Stéphane Galland, Christophe Nicolle, Igor Tchappi Haman, Ansar-Ul-Haque Yasar, and Kary Främling,
Between the Megalopolis and the Deep Blue Sky: Challenges of Transport with UAVs in Future Smart Cities,
International Conference on Autonomous Agents and Multiagent Systems (CORE Rank: **A***), Montreal, Canada, pp. 1649-1653 (2019). URL: http://www.ifaamas.org/Proceedings/aamas2019/pdfs/p1649.pdf

- Yazan Mualla, Igor Tchappi Haman, Amro Najjar, Timotheus Kampik, Stéphane Galland, and Christophe Nicolle,
Human-Agent Explainability: An Experimental Case Study on the Filtering of Explanations,
12th International Conference on Agents and Artificial Intelligence (ICAART), Volume 1: HAMT, ISBN 978-989-758-395-7, ISSN 2184-433X, pages 378-385 (2020) [presented by me in Valletta, Malta]. DOI: 10.5220/0009382903780385

International Workshop with Proceedings

- Yazan Mualla, Wenshuai Bai, Stéphane Galland, and Christophe Nicolle,
Comparison of Agent-Based Simulation Frameworks for Unmanned Aerial Transportation Applications,
Procedia Computer Science 130, Elsevier, pp. 791-796 (2018) [ABMTRANS workshop, presented by me in Porto, Portugal]. DOI: 10.1016/j.procs.2018.04.137

- Yazan Mualla, Timotheus Kampik, Igor Tchappi Haman, Amro Najjar, Stéphane Galland, and Christophe Nicolle,

Explainable Agents as Static Web Pages: UAV Simulation Example,
2nd International Workshop on EXplainable TRansparent Autonomous Agents and Multi-Agent Systems, AAMAS, Auckland, New Zealand, pp. 149-154 (2020) [presented by me online]. URL: https://link.springer.com/chapter/10.1007/978-3-030-51924-7_9

- Davide Calvaresi, Yazan Mualla, Amro Najjar, Stéphane Galland, and Michael Schumacher,
Explainable Multi-Agent Systems through Blockchain Technology,
1st International workshop on eXplainable TRansparent Autonomous Agents and Multi-Agent Systems, AAMAS, Montreal, Canada, pp. 41-58, Springer, Cham (2019). URL: https://link.springer.com/chapter/10.1007/978-3-030-30391-4_3

International Conference/Workshop without Proceedings

- Yazan Mualla, Amro Najjar, Timotheus Kampik, Igor Tchappi Haman, Stéphane Galland, and Christophe Nicolle,
Towards Explainability for a Civilian UAV Fleet Management Using an Agent-Based Approach,
1st Workshop on Explainable AI in Automated Driving: A User-Centered Interaction Approach, Utrecht, Netherland, arXiv preprint arXiv:1909.10090 (2019). URL: https://arxiv.org/abs/1909.10090

- Alexandre Lombard, Yazan Mualla, Stéphane Galland, and Jocelyn Buisson,
Software Architecture for Drone Simulation in 3D,
European Forum for the SARL Users and Developers, Leuven, Belgium (2019)

- Jocelyn Buisson, Yazan Mualla, Alexandre Lombard, and Stéphane Galland,
Traffic Simulation with Sarl,
European Forum for the SARL Users and Developers, (2019) [presented by me in Leuven, Belgium]

- Stéphane Galland, Yazan Mualla, Hui Zhao, Sebastian Rodriguez, Amro Najjar, and Nicolas Gaud,
Model Transformations from the SARL Agent-Oriented Programming Language to an Object-Oriented Programming Language,
Baroglio, Cristina and Hübner, Jomi Fred and Winikoff, Michael, Auckland, New Zealand, Special IJ-AOSE track of the 8th International Workshop on Engineering Multi-Agent Systems (EMAS 2020)

F.2/ Other Publications

Number of other publications	
International journal with reading committee	1
International conference with proceedings	3
International workshop with proceedings	4
National conference with proceedings	4
Total	**12**

International Journal with Reading Committee

- Igor Tchappi Haman, Stéphane Galland, Vivient Corneille Kamla, Jean-Claude Kamgang, Yazan Mualla, Amro Najjar, and Vincent Hilaire,
A Critical Review of Holonic Technology in Traffic and Transportation Fields,
International Journal of Engineering Applications of Artificial Intelligence (SJR Quartile: **Q1**, Clarivate Analytics Impact Factor 2019: **4.201**), vol. 90, pp. 1-54 (2020). DOI: 10.1016/j.engappai.2020.103503

International Conference with Proceedings

- Yazan Mualla, Robin Vanet, Amro Najjar, Olivier Boissier, and Stéphane Galland,
AgentOil: A Multi-Agent-Based Simulation of the Drilling Process in Oilfields,
16th International Conference on Practical Applications of Agents and Multi-Agent Systems, Lecture Notes in Artificial Intelligence, Springer, pp. 339-343. Springer, Cham (2018) [presented by me in Toledo, Spain]. URL: https://link.springer.com/chapter/10.1007/978-3-319-94580-4_34

- Hui Zhao, Yazan Mualla, Stéphane Galland, Igor Tchappi Haman, Tom Bellemans, and Ansar-Ul-Haque Yasar,
Decision-Making under Time Pressure when Rescheduling Daily Activities,
11th International Conference on Ambient Systems, Networks and Technologies (ANT), Warsaw, Poland, Procedia Computer Science 170, pp. 281-288 (2020). URL: https://doi.org/10.1016/j.procs.2020.03.041

- Emmanuel Dimitry Ngounou Ntoukam, Vivient Corneille Kamla, Yazan Mualla, Stéphane Galland, Jean-Claude Kamgang, and Yves Sébastien Emvudu Wono,
A Multi-Agent Model of the Population Dynamics of Mirids in a Cocoa Farm,
1st Conference of the Cameroon Academy of Young Scientists, Yaoundé, Cameroon (2019).

National Conference with Proceedings

- Yazan Mualla, Igor Tchappi Haman, Amro Najjar, Stéphane Galland, Robin Vanet, and Olivier Boissier,
  Modélisation multi-agent des opérations semi-autonomes dans un système cyber-physique de forage pétrolier ou gazier,
  27e Journées Francophones sur les Systèmes Multiagents, Association Française d'Intelligence Artificielle, Toulouse, France, pp. 126-135 (2019). URL: https://team.inria.fr/chroma/files/2019/07/Actes_CH_PFIA2019-1.pdf

- Igor Tchappi Haman, Stéphane Galland, Yazan Mualla, Amro Najjar, Vivient Corneille Kamla, and Jean-Claude Kamgang,
  Modèle dynamique et multiniveau holonique basé sur la densité : application au trafic routier à grande échelle,
  27e Journées Francophones sur les Systèmes Multiagents, Association Française d'Intelligence Artificielle, Toulouse, France, pp. 136-145 (2019). URL: https://team.inria.fr/chroma/files/2019/07/Actes_CH_PFIA2019-1.pdf

- Amro Najjar, Yazan Mualla, Gauthier Picard, and Kamal Singh,
  Négociation multi-agent one-to-many et mécanismes de coordination pour la gestion de la satisfaction des utilisateurs d'un service,
  26e Journées Francophones sur les Systèmes Multi-Agents, Métabief, France, pp. 95-104 (2018). URL: https://hal.archives-ouvertes.fr/hal-01861569/

- Emmanuel Dimitry Ngounou Ntoukam, Vivient Corneille Kamla, Yazan Mualla, Igor Tchappi Haman, Stéphane Galland, Jean-Claude Kamgang, and Yves Sébastien Emvudu Wono,
  Towards a Multi-Agent Model to Prevent Damage Caused by Cocoa Mirids to Cocoa Pods,
  17e Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA), Toulouse, France, pp. 10-17 (2019). URL: https://jeannicod.ccsd.cnrs.fr/UNIV-BM/hal-02160273v1

International Workshop with Proceedings

- Yazan Mualla, Amro Najjar, Olivier Boissier, Stéphane Galland, Igor Tchappi Haman, and Robin Vanet,
  A Cyber-Physical System for Semi-Autonomous Oil&Gas Drilling Operations,
  5th Workshop on Collaboration of Humans, Agents, Robots, Machines and Sensors, IEEE Computer Society, pp. 514-519 (2019) [presented by me in Naples, Italy]. DOI: 10.1109/IRC.2019.00107

- Yazan Mualla, Amro Najjar, Robin Vanet, Olivier Boissier, and Stéphane Galland,
  Towards a Real-Time Mitigation of High Temperature while Drilling Using a Multi-Agent System,
  International Workshop on Real-Time compliant Multi-Agent Systems, Stockholm, Sweden, pp. 77-92 (2018). URL: http://ceur-ws.org/Vol-2156/

- Amro Najjar, Yazan Mualla, Gauthier Picard, and Kamal Singh,
  One-to-Many Multi-Agent Negotiation and Coordination Mechanisms to Manage User Satisfaction,
  11th International Workshop on Automated Negotiations, Stockholm, Sweden (2018). URL: https://hal.archives-ouvertes.fr/hal-01802513/

- Emmanuel Dimitry Ngounou Ntoukam, Jean-Claude Kamgang, Vivient Corneille Kamla, Yazan Mualla, Igor Tchappi Haman, Stéphane Galland, and Yves Sébastien Emvudu Wono,
  Agent-Based Model of Cocoa Mirids at the Scale of a Cocoa Farm,
  4th International Workshop on Agent-based Modeling and Applications with SARL (SARL-20), Procedia Computer Science 170, pp.1180-1185 (2020). URL: https://doi.org/10.1016/j.procs.2020.03.032

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — https://github.com/gallandarakhneorg/tex-templates
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/

école doctorale **sciences pour l'ingénieur et microtechniques**

Title: Explaining the Behavior of Remote Robots to Humans: An Agent-based Approach

Keywords: Explainable Artificial Intelligence, Multi-agent Systems, Human-Computer Interaction.

Abstract:

With the widespread use of Artificial Intelligence (AI) systems, understanding the behavior of intelligent agents and robots is crucial to guarantee smooth human-agent collaboration since it is not straightforward for humans to understand the agent's state of mind. Recent studies in the goal-driven eXplainable AI (XAI) domain have confirmed that explaining the agent's behavior to humans fosters the latter's understandability of the agent and increases its acceptability. However, providing overwhelming or unnecessary information may also confuse human users and cause misunderstandings. For these reasons, the parsimony of explanations has been outlined as one of the key features facilitating successful human-agent interaction with a parsimonious explanation defined as the simplest explanation that describes the situation adequately. While the parsimony of explanations is receiving growing attention in the literature, most of the works are carried out only conceptually.

This thesis proposes, using a rigorous research methodology, a mechanism for parsimonious XAI that strikes a balance between simplicity and adequacy. In particular, it introduces a context-aware and adaptive process of explanation formulation and proposes a Human-Agent Explainability Architecture (HAExA) allowing to make this process operational for remote robots represented as Belief-Desire-Intention agents. To provide parsimonious explanations, HAExA relies first on generating normal and contrastive explanations and second on updating and filtering them before communicating them to the human.
To evaluate the proposed architecture, we design and conduct empirical human-computer interaction studies employing agent-based simulation. The studies rely on well-established XAI metrics to estimate how understood and satisfactory the explanations provided by HAExA are. The results are properly analyzed and validated using parametric and non-parametric statistical testing.

Titre : Expliquer le comportement de robots distants à des utilisateurs humains : une approche orientiée-agent

Mots-clés : Intelligence artificielle explicable, Systèmes multi-agents, Interaction homme-machine.

Résumé :

Avec l'émergence et la généralisation des systèmes d'intelligence artificielle, comprendre le comportement des agents artificiels, ou robots intelligents, devient essentiel pour garantir une collaboration fluide entre l'homme et ces agents. En effet, il n'est pas simple pour les humains de comprendre les processus qui ont amenés aux décisions des agents. De récentes études dans le domaine l'intelligence artificielle explicable, particulièrement sur les modèles utilisant des objectifs, ont confirmé qu'expliquer le comportement d'un agent à un humain favorise la compréhensibilité de l'agent par ce dernier et augmente son acceptabilité. Cependant, fournir des informations trop nombreuses ou inutiles peut également semer la confusion chez les utilisateurs humains et provoquer des malentendus. Pour ces raisons, la parcimonie des explications a été présentée comme l'une des principales caractéristiques facilitant une interaction réussie entre l'homme et l'agent. Une explication parcimonieuse est définie comme l'explication la plus simple et décrivant la situation de manière adéquate. Si la parcimonie des explications fait l'objet d'une attention croissante dans la littérature, la plupart des travaux ne sont réalisés que de manière conceptuelle.

Dans le cadre d'une méthodologie de recherche rigoureuse, cette thèse propose un mécanisme permettant d'expliquer le comportement d'une intelligence artificielle de manière parcimonieuse afin de trouver un équilibre entre simplicité et adéquation. En particulier, il introduit un processus de formulation des explications, sensible au contexte et adaptatif, et propose une architecture permettant d'expliquer les comportements des agents à des humains (HAExA). Cette architecture permet de rendre ce processus opérationnel pour des robots distants représentés comme des agents utilisant une architecture de type Croyance-Désir-Intention. Pour fournir des explications parcimonieuses, HAExA s'appuie d'abord sur la génération d'explications normales et contrastées, et ensuite sur leur mise à jour et leur filtrage avant de les communiquer à l'humain.
Nous validons nos propositions en concevant et menant des études empiriques d'interaction homme-machine utilisant la simulation orientée-agent. Nos études reposent sur des mesures bien établies pour estimer la compréhension et la satisfaction des explications fournies par HAExA. Les résultats sont analysés et validés à l'aide de tests statistiques paramétriques et non paramétriques.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France