Explainable and Ethical AI: A Perspective on Argumentation and Logic Programming

Roberta Calegari[1](✉), Andrea Omicini[2], and Giovanni Sartor[1]

[1] Alma AI – Alma Mater Research Institute for Human-Centered Artificial Intelligence, Alma Mater Studiorum—Università di Bologna, Bologna, Italy
roberta.calegari@unibo.it

[2] Dipartimento di Informatica – Scienza e Ingegneria (DISI), Alma Mater Studiorum—Università di Bologna, Cesena, Italy

Abstract. In this paper we sketch a vision of explainability of intelligent systems as a logic approach suitable to be injected into and exploited by the system actors once integrated with sub-symbolic techniques.

In particular, we show how argumentation could be combined with different extensions of logic programming – namely, abduction, inductive logic programming, and probabilistic logic programming – to address the issues of explainable AI as well as some ethical concerns about AI.

Keywords: Explainable AI · Ethical AI · Argumentation · Logic programming · Abduction · Probabilistic LP · Inductive LP

1 Introduction

In the new artificial intelligence (AI) era, intelligent systems are increasingly relying on sub-symbolic techniques such as deep learning [2,7]. Since opaqueness of most sub-symbolic techniques engenders fear and distrust, the behaviour of intelligent systems should be observable, explainable, and accountable—which is the goal of the eXplainable Artificial Intelligence (XAI) field [2,8].

In this paper we focus on logic-based approaches and discuss their potential to address XAI issues especially in pervasive scenarios that can be designed as open multi-agent systems (MAS)—the reference for the design of intelligent systems [7,44,45]. In particular, this paper proposes an architecture for delivering (ubiquitous) symbolic intelligence to achieve explainability in pervasive contexts based on two assumptions: *(i) ubiquitous symbolic intelligence* is the key to making the environment truly smart and self-explainable, *(ii) declarativeness* and *transparency* can lead to the injection of ethical behaviours—see e.g. [34].

Roberta Calegari and Giovanni Sartor have been supported by the H2020 ERC Project "CompuLaw" (G.A. 833647). Andrea Omicini has been supported by the H2020 Project "AI4EU" (G.A. 825619).

M. Baldoni and S. Bandini (Eds.): AIxIA 2020, LNAI 12414, pp. 19–36, 2021.
https://doi.org/10.1007/978-3-030-77091-4_2

The architecture enables on-demand symbolic intelligence injection only *where* and *when* required. Sub-symbolic techniques – e.g., deep networks algorithms – are therefore part of our vision and can coexist in the system even in case they are not fully explainable. One of the main requirements of any system is then to identify which parts need to be explained – for ethical or legal purposes, responsibility issues, etc. – and which ones can instead remain opaque.

Logic-based approaches already play a well-understood role in the engineering of intelligent (multi-agent) systems; declarative, logic-based approaches have the potential to represent an alternative way of delivering symbolic intelligence, complementary to the one pursued by sub-symbolic approaches [7]. Logic-based technologies address opaqueness issues, and, once suitably integrated with argumentation capabilities, can provide for features like interpretability, observability, accountability, and explainability. In our vision, explainability depends on a system's capability of conversing and debating about situations and choices, providing reports and insights into what is happening. An explanation can be seen as a sort of *conversation* among the person asking for clarification and the system actors – agents, environment, and (e-)institution. As far as ethics is concerned, LP has recently been deeply studied by the research community, precisely in relation to the implementation of ethical machines and systems [40].

Argumentation is the spearhead of the proposed approach, yet – in order to tackle the AI requirements for ubiquitous intelligence – it should be strongly intertwined with logic programming and its extensions. In particular, our vision of symbolic intelligence leverages argumentation, abduction, inductive logic programming, and probabilistic logic programming, along the line of some recent research works—e.g., [22,26,31].

2 Explanation: Meaning and Roles

The first issue to be clarified when it comes to explainability is the acceptation of the term. Since explainability has become one of the hottest research topics in AI, the very notion of explainability has become the subject of scientific debate, also aimed at defining related concepts and terms such as explanation, interpretability, and understandability. Yet, there is still no widely-shared definition, also due to the pervasiveness of terms from the common language [32].

Formally defining those terms is not the goal of this paper. Instead, in the following, we point out some issues to consider when dealing with explainability and discuss how they affect the design of the proposed architecture.

Explanation vs Interpretation. The terms "interpretability" and "explainability" are often used carelessly and interchanged in the context of XAI [16]. Although the two terms are closely related and both contributing to the ultimate goal of understandability, they should be kept well distinct. Accordingly, here we borrow the definition of *interpretation* from logic, where the word essentially describes the operation of binding objects to their actual meaning in some context—thus, the goal of interpretability is to convey to humans the meaning of data [15]. Then, we conceive *explanation* can be seen as an activity of symbolic representation and transformation by the *explanator*, aimed at making the subjective activity of interpretation by the *explainee* easier [16,32].

Once this distinction has been made, two general remarks are useful: *(i)* most XAI approaches proposed in the recent literature mostly focus on interpretability, *(ii)* to reach explainability, a mechanism for unwinding a reasoning and a corresponding conversation enabler is somehow required. This mechanism allows distinct actors' roles to be taken into account in the explanation process. Also, conversations can be used to clarify non-understandable explanations.

Explanation Actors and Kind of Explanation. Who are the explanators, and who the explainees? Can we proceed beyond the simplistic hypothesis that software systems/agents are the explanators, and explainees are just humans? In the literature [1,24,39] explanators are typically software components (i.e., agents in multi-agent systems), explainees are intended to be humans.

Instead, every possible direction that explanation could follow in AI systems should be explored [32]: human-agent, agent-human, agent-agent. Accordingly, tools and methods for explainability should be light-weight, easy integrable in existing technologies, embeddable in different AI techniques, and easily usable by software developers and engineers. For the same reasons, interoperability is one of the main requirements we take into account in the design of the architecture.

Furthermore, the sort of explanation that a system can provide is relevant, indeed. An explanation for a human is not necessarily useful for an agent, and viceversa. So, for instance, a logic-based tool generating an effective explanation for agents may not necessarily be immediately useful for humans; yet, once integrated with the appropriate AI techniques – such as tools for translation into natural language – it could become effective for humans, too.

3 Logic Techniques for XAI

3.1 Why Logic?

Our driving question here is: "What is or can be the added value of logic programming for implementing machine ethics and explainable AI?" The main answer lies in the three main features of LP: *(i)* being a declarative paradigm, *(ii)* working as a tool for knowledge representation, and *(iii)* allowing for different forms of reasoning and inference. These features lead to some properties for intelligent systems that can be critical in the design of ubiquitous intelligence.

Provability. By relying on LP, the models can provide for well-founded semantics ensuring some fundamental computational properties – such as correctness and completeness. Moreover, extensions can be formalised, well-founded as well, based on recognised theorems—like for instance, correctness of transitive closure, strongly equivalent transformation, modularity, and splitting set theorem. Provability is a key feature in the case of trusted and safe systems.

Explainability. The explainability feature is somehow intrinsic in LP techniques. Formal methods for argumentation-, justification-, and counterfactual-based methods are often based on a logic programming approach [21,35,40]. These techniques make the system capable to engage in dialogues with other actors to communicate its reasoning, explain its choices, or to coordinate in the pursuit of a common goal. So, the explanation can be a dialogue showing insights on reasoning or, again, the explanation can be the unraveling of causal reasoning based on counterfactual. Counterfactuals are the base for hypothetical reasoning, a necessary feature both for explanation and machine ethics. Furthermore, other logical forms of explanation can be envisaged via non-monotonic reasoning and argumentation, through a direct extension of the semantics of LP.

Expressivity and Situatedness. As far as the knowledge representation is concerned, the logical paradigm brings non-obvious advantages—beyond the fact of being human-readable. First of all, a logical framework makes it possible to grasp different nuances according to the extensions considered—e.g., nondeterminism, constraints, aggregates [20]. Also, assumptions and exceptions can be made explicit, as well as preferences—e.g., weighted weak constraints [4]. Finally, extensions targeting the Internet of Things can allow knowledge to be situated in order to be able to capture the specificities of the context in which it is located [12]. Expressive, flexible, and situated frameworks are needed to cover various problems and reasoning tasks closely related to each other.

Hybridization. One of the strengths of computational logic is to make it possible the integration of diverse techniques [11,42]—e.g., logic programming paradigms, database technologies, knowledge representation, non-monotonic reasoning, constraint programming, mathematical programming, etc. This makes it possible to represent the heterogeneity of the contexts of intelligent systems – also in relation to the application domains – and to customise as needed the symbolic intelligence that is provided while remaining within a well-founded formal framework.

3.2 User Requirements for XAI

Before we move into the discussion of the main extensions that a symbolic intelligence engine needs to have in order to inject explainability, let us define what we should expect from an explainable system and what kind of intelligence the system is supposed to deal with.

- $\mathbf{R_1}$ First of all, the system should be able to answer *what* questions, i.e., it should provide query answering and activity planning in order to achieve a user-specified goal.
- $\mathbf{R_2}$ The system should be able to answer *why* questions, i.e., it should provide explanation generation (in the form of text, images, narration, conversation) and diagnostic reasoning.
- $\mathbf{R_3}$ The system should be able to answer *what if* questions, i.e., it should provide counterfactual reasoning and predictions about what would happen under certain conditions and given certain choices.

- $\mathbf{R_4}$ The system should be able to answer *which* questions, i.e., it should be able to choose which scenarios to implement, once plausible scenarios have been identified as in the previous point. The choice should result from the system's preferences, which could possibly be user-defined or related to the context.
- $\mathbf{R_5}$ The system should be able to provide *suggestions*, i.e., to indicate what to do given the current state of affairs, exploiting hypothetical reasoning.
- $\mathbf{R_6}$ The system should be able to support two types of intelligence and therefore reasoning, i.e., *reactive reasoning* – related to the data and the current situation – and *deliberative reasoning*—related more to consciousness, knowledge, and moral, normative principles.

Even if only $\mathbf{R_2}$ is strictly and explicitly related to the explainability feature, also the other requirements can help to understand and interpret the system model, so all the above-mentioned requirements can be identified as mandatory for reaching ethical features such as interpretability, explainability, and trustworthiness. According to the requirements, in the following we discuss what logical approach should be part of an engine that enables symbolic intelligence to be injected in contexts demanding the aforementioned properties.

3.3 Logic Approaches and Technologies Involved for XAI

In our vision, logic programming is the foundation upon which the architecture for a symbolic intelligence engine can be built, enabling an intelligent system to meet the $\mathbf{R_1}$ requirement. Clearly, enabling different forms of inference and reasoning – e.g., non-monotonic reasoning – paves the way for the possibility to get different answers (appropriate to the context) to the *what* questions. Furthermore, the techniques of inference and reasoning grafted into the symbolic engine make it possible to reason about preferences by meeting requirement $\mathbf{R_4}$.

However, LP needs to be extended in order to address explainability in different AI technologies and applications, and to be able to reconcile the two aspects of intelligence present in today's AI systems—namely, *reactive* and *deliberative* reasoning. In particular, in the following we show how argumentation, abduction, induction, and probabilistic LP can be fundamental ingredients to shape explainable and ethical AI.

Argumentation. In this vision, argumentation is the enabler to meet requirement $\mathbf{R_2}$. Argumentation is a required feature of the envisioned symbolic intelligence engine to enable system actors to talk and discuss in order to explain and justify judgments and choices, and reach agreements.

Several existing works set the maturity of argumentation models as a key enabler of our vision [25,30]. In spite of the long history of research in argumentation and the many fundamental results achieved, much effort is still needed to effectively exploit argumentation in our envisioned framework. First, research on formal argumentation has mostly been theoretical: practical applications to real-world scenarios have only recently gained attention, and are not yet reified in a ready-to-use technology [10]. Second, many open issues of existing argumentation frameworks concern their integration with contingency situations and situated reasoning to achieve a blended integration of reactive and deliberative reasoning. Finally, the argumentation architecture should be designed in order to be highly scalable, distributed, open, and dynamic, and hybrid approaches should be investigated.

Abduction. Abduction is the enabling technique to meet $\mathbf{R_3}$. Abduction, in fact, allows plausible scenarios to be generated under certain conditions, and enables hypothetical reasoning, including the consideration of counterfactual scenarios about the past. Counterfactual reasoning suggests thoughts about what might have been, what might have happened if any event had been different in the past. What if I have to do it today? What have I learned from the past? It gives hints about the future by allowing for the comparison of different alternatives inferred from the changes in the past. It supports a justification of why different alternatives would have been worse or not better. After excluding those abducibles that have been ruled out a priori by integrity constraints, the consequences of the considered abducibles have first to be evaluated to determine what solution affords the greater good. Thus, reasoning over preferences becomes possible. Counterfactual reasoning is increasingly used in a variety of AI applications, and especially in XAI [23].

Probabilistic Logic Programming. Probabilistic logic programming (PLP) allows symbolic reasoning to be enriched with degrees of uncertainty. Uncertainty can be related to facts, events, scenarios, arguments, opinions, and so on. On the one side, PLP allows abduction to take scenario uncertainty measures into account [37]. On the other side, probabilistic argumentation can account for diverse types of uncertainty, in particular uncertainty on the credibility of the premises, uncertainty about which arguments to consider, and uncertainty on the acceptance status of arguments or statements [38]. Reasoning by taking into account probability is one of the key factors that allow a system to fully meet $\mathbf{R_4}$ and $\mathbf{R_5}$, managing to formulate well-founded reasoning on which scenario to prefer and which suggestions to provide as outcomes.

Inductive Logic Programming. Inductive logic programming (ILP) can help us bridging the gap between the symbolic and the sub-symbolic models—by inserting data and context into the reasoning. As already expressed by $\mathbf{R_6}$, data, context, and reactive reasoning are key features to take into account when designing intelligence. ILP makes it possible to learn from data enabling inductive construction of first-order clausal theories from examples and background knowledge. ILP is a good candidate to meet $\mathbf{R_6}$ and preliminary studies show ILP can be the glue between symbolic techniques and sub-symbolic ones such as numerical/statistical machine learning (ML) and deep learning [3].

All these techniques should be suitably integrated into a unique consistent framework, in order to be used appropriately when needed: they should be involved in the engineering of systems and services for XAI.

4 System Architecture

Figure 1 summarises our vision by highlighting the main roles involved in the system as well as the main activity flows. The grey boxes represent the technologies involved in the vision, while arrows represent the expected provided functionalities. The symbolic reasoner embodies the unique framework integrating the aforementioned logic approaches.

On one side, knowledge is collected from various sources – e.g., domain-specific knowledge, ontologies, sensors raw data – and is then exploited by agents that live in a normative environment. Note that we mean to exploit already existing techniques to convert ML knowledge into logic KB [9] and to explore other possibilities – always related to the exploitation of the aforementioned LP approaches – to explain (part of) deep knowledge.

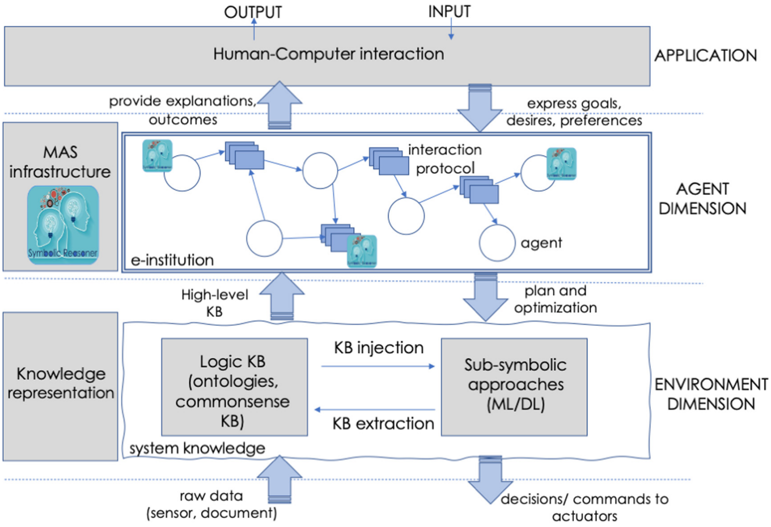

Fig. 1. Main architecture components and techniques in the vision.

The cognitive ability of the system is expanded with the concept of symbolic (micro-)intelligence which provides the techniques of symbolic reasoning discussed in Sect. 3 and tailored to LP. The multi-agent system, also thanks to its rational reasoning and argumentation capabilities, can provide outcomes to the users as well as explanations for their behaviours. On the other side, humans can insert input into the system – like desires, preferences, or goals to achieve – and these are transposed into agents' goals, corresponding activity planning, and lower-level commands for actuators.

Our vision stems from two basic premises: *(i)* knowledge is locally scattered in a distributed environment, hence its situated nature; *(ii)* symbolic capabilities are available over this knowledge, with the goal of extending local knowledge through argumentation, induction, deduction, abduction, and probabilistic reasoning and therefore pave the way towards explanations generation; *(iii)* distributed knowledge can be considered as compartmentalised in distinct knowledge modules and can be used by itself, or by referring to other modules for specific questions (according to the model of modular LP).

4.1 Main Enabling Technologies

The architecture described so far is rooted in a well-founded integration of different AI approaches and techniques. In the following, we describe the main ones – as well as open challenges and issues – well aware that a methodology for their integration is far from being already defined.

Knowledge Representation and Sharing. Knowledge representation and related techniques are some of the main ingredients of the envisioned distributed system, to enable conversation, argumentation, and reasoning. The system knowledge has to take into account domain-specific knowledge and large-scale ontologies as repositories to interpret the knowledge bases available to the agents and to reason and argument over it. Knowledge could be continuously modified, adapted, and refined by the agents, according to their experience and perception of the environment or to learning from experience. Accordingly, the knowledge base is plausible that is assembled by two main sources: on the one hand, ontologies and hand-crafted rules, on the other hand, rules learned from big data. Advances in machine learning will allow extracting knowledge from this data and merging it with the former. Hybrid approaches dealing with the integration of symbolic and sub-symbolic approaches become of paramount importance.

In this context, there are several issues and challenges to be tackled, to cite a few, automatic extraction of knowledge from ML models, extraction of commonsense knowledge from the context, integration of the diverse knowledge in an appropriate logical language that allows argumentation and inference process to be performed. Several research fields are already facing these issues, but the general problem is far from being solved. For sure, we believe that a suitable integration of symbolic and sub-symbolic approaches can help in the achievement of the construction of proper system knowledge. In addition, agents' mental state, including cultural features and commonsense knowledge, is necessary to deal with humans and be on par with their knowledge of real-world concepts. Moreover, their emotional state, including the support of trust and the capability to entertain the user in a believable way is fundamental. To this end, possibly new forms of knowledge representation should be envisioned and synergistically integrated enabling argumentation and semantic reasoning over it.

Finally, it is worth emphasising that the sharing of knowledge, therefore the possibility of making it explicit – and so explainable –, is one of the main purposes of XAI, as well as of human beings.

Machine Learning. In our vision, a fundamental role is played by machine learning involved in different phases—namely, data processing & rule learning, and planning.

Data Processing and Rule Learning. At the most straightforward level, machine learning techniques are clearly involved in raw input data elaboration, coming from sensors and/or documents, into more complex, high-level, structured information. Moreover, agents should be able to learn policies from past experience, by adapting both to the changing environment, and to the continuous progress of the society. Data aggregation, feature extraction, clustering, classification, data, and pattern mining techniques are typically employed today to reach these objectives. We believe that hybrid approaches could provide promising solutions to these tasks, by merging logic with probabilistic models and statistical learning, so to efficiently handle advantages of both symbolic and sub-symbolic approaches and moving towards explainable systems [9]. As highlighted above, the ML knowledge should somehow be translated into logical knowledge and properly merged with logical knowledge coming from ontologies or domain-expert norm translation or similar.

Planning. Distributed problem solving, planning, reinforcement learning, and cooperation [41] are some of the well-known ML techniques exploited in MAS. Our framework adds the challenge of integrating these techniques in the argumentation setting so that the planning and cooperation derive from a continuous, natural interaction between agents with the environment. Once the user has specified his desires, the agent must be able to achieve them, interacting and coordinating with other individuals and with the e-institution to define the actions to perform and consequently defining appropriate plans to reify the decisions.

Symbolic Reasoning Engine. The symbolic reasoning engine is the cornerstone of the proposed approach. Each of the symbolic techniques described in Sect. 3 allows one of the XAI requirements to be achieved. Accordingly, the foundation of our vision is to have a symbolic reasoning engine – which carries out the techniques discussed above – to be injectable on-demand into the various system's components—agents and/or environment and/or institutions. Symbolic (micro-)intelligence architecture [5,33] is exploited to deliver symbolic intelligence according to the new paradigms of AI. The architecture of symbolic (micro-)intelligence should enable – only where and when necessary – actions at the micro-level, to respond to local and specific needs [6]. Symbolic (micro-)intelligence complements agents' cognitive processes because it augments the cognitive capabilities of agents, by embodying situated knowledge about the local environment along with the relative inference processes, based on argumentation, abduction, ILP, and PLP. Beyond the open issues that belong to each of the logical fields considered, the main open issue to be tackled is their integration within the desired framework—i.e., a well-founded integration methodology. The methodology should employ a multidisciplinary approach that combines expertise in the fields of software engineering, natural language processing, argumentation, logic, ontologies, and of course distributed and autonomous systems.

MAS and Normative MAS: Middleware. From a more implementation-oriented perspective, given that conversations are a new means of orchestrating the activities of distributed agents, an open research question – and a key one, too – is to understand which services should a middleware provide in order to support such distributed conversations [13].

The multi-agent infrastructure needs not only to enable coordination among system actors but also include the possibility of customisable and reactive artefacts, capable of incorporating regulation and norms and micro-intelligence. Moreover, the middleware should provide support for discussions via an open and shared discussion space, enabling dialogue among components that do not necessarily know each other in advance, and also providing services and or techniques for sharing knowledge, e.g., a tuple space [28]. However, unlike traditional tuple space models, the evolution of the conversation, the argumentation process, and the reached consensus should be taken into account, also to be exploited in similar situations and/or to provide explanations. The best way to build such shared dialogue spaces – also taking into account different sources of knowledge (e.g., commonsense kb) and different artefacts acting as both law enforcers and intelligence promoters – is a fertile ground for research.

Human-Computer Interaction. Knowledge sharing, also in terms of explanation, requires a form of conversation among agents and humans, but also agents and agents, and conversation requires mind-reading – in terms of ability to understand motivations, beliefs, goals of others –, or, more generally, a theory of mind [43]. Accordingly, for human interaction, techniques coming from natural language processing, computer vision speech recognition become essential components of our vision.

The challenge here is twofold. On the one hand, the challenge is always related to the distributed issues, i.e., making commands possibly understandable to a multitude of agents and vice-versa. Existing algorithms should, therefore, be adapted for dealing with distributed and pervasive environments. On the other hand, existing techniques should be enhanced to understand hidden emotions possibly based on human culture, tone of voice and so on.

5 Preliminary Investigation: Examples

To ground our proposal, let us discuss a preliminary example from a case study in the area of traffic management, considering the near future of self-driving cars. In that scenario, cars are capable of communicating with each other and with the road infrastructure while cities and roads are suitably enriched with sensors and virtual traffic signs able to dynamically interact with cars to provide for information and supervision.

Accordingly, self-driving cars need to *(i)* exhibit some degree of intelligence for taking autonomous decisions; they need to *(ii)* converse with the context that surrounds them, *(iii)* have humans in the loop, *(iv)* respond to the legal setting characterising the environment and the society, and *(v)* offer explanations when required—e.g., in case of accidents to determine causes and responsibilities. Figure 2 (left) contains a possible example of the logical knowledge that, despite its simplicity, highlights the main different sources of knowledge taken into account in such a scenario. First of all, knowledge includes data collected by vehicle sensors as well as the beliefs of vehicles—possibly related to the outcome of a joint discussion among other entities in the system. Then, commonsense rules enrich the system knowledge, for instance, linking perceptions to beliefs about the factual situations at stake. Also, commonsense rules can state general superiority relations, such as that sensors' perceptions must be considered prevailing over vehicles' beliefs. An additional source of knowledge is e-institution knowledge. Loosely speaking, e-institutions are computational realisations of traditional institutions that incarnate the global system norms as global, national, state, and local laws, regulations, and policies. For instance, the e-institution knowledge defined in Fig. 2 declares that the general speed limit – according to German federal government – is 100 km/h outside built-up areas (no highways). In addition, a general norm is stated by the e-institution declaring that the overtake is permitted only if it is not raining. Another possible source of knowledge is situated knowledge collected by the surrounding context (infrastructure) that can include specific local rules stating exceptions to the general e-institutions rules. For instance, in the example, situated knowledge states that in the road being represented the general speed limit only applies if it does not rain, otherwise vehicles must slow down to 60 km/h. Note that in the example we list all the different kinds of knowledge in a unique file, but a suitable technology that embodies the envisioned architecture needs to manage different modules and to combine them—depending on the situation.

Figure 2 (right) shows some system outcomes, depending on the situation. All examples have been implemented and tested on the preliminary implementation of the system—namely, Arg-tuProlog (Arg2P in short) [36][1]. Arg2P – designed according to the vision discussed in this paper – is a lightweight modular argumentation tool that fruitfully combines modular logic programming and legal reasoning according to an argumentation labelling semantics where any statement (and argument) is associated with one label that is IN, OUT, UND, respectively meaning that the argument is accepted, rejected, or undecided. Example 1 is run without taking into account the superiority relation of perceptions over beliefs. In this situation, beliefs and perceptions are in conflict and no decision can be taken by the system, i.e., vehicles can base their decision only by taking into account the e-institution obligation and cannot be sure on the permission of overtaking. Example 2, instead, takes superiority relation into account, and according to the fact that sensor perception imposes a speed limit of 60 km/h and negate permission to overtake. The argumentation process among the system actors makes them meet the conclusion that it rains, so both vehicles, despite their beliefs, will set the maximum speed to 60 km/h. Conversely, Example 3 is run by negating rain perception. The system then recognises that it is not raining, so vehicle speed can be set to 100 km/h, and overtakes are allowed.

[1] http://arg2p.apice.unibo.it.

```
%******** SYSTEM KB *******************
%******** ************ ***************

%** PERCEPTIONS and BELIEFS **
pr1:[] => perception(rain).

b1: [] => belief(agent1, rain).
b2: [] => -belief(agent2, rain).


%** GENERAL-COMMONSENSE KB **
% perceptions/beliefs translation
r1: perception(X) => fact(X).
r2: -perception(X) => -fact(X).

r3: belief(A, X) => fact(X).
r4: -belief(A, X) => -fact(X).

%** GENERAL-COMMONSENSE KB **
% perceptions are superior to beliefs
sup(r1,r3).
sup(r1,r4).
sup(r2,r3).
sup(r2,r4).


%** e-INSTITUTION RULES **
% permissions and obligations
o1: [] => o(max_speed(100)).
p1: -fact(rain) => p(overtaking).

%** SITUATED LOCAL KB **
% specific road obligation
% if rains max speed 60 km/h
r5: fact(rain) => speed(60).
r6:-fact(rain),o(max_speed(X))=>speed(X).
```

```
%**** Example 1 ****
IN(accepted) =======>
[obl, [max_speed(100)]]
[neg, belief(agent2, rain)]
[belief(agent1, rain)]
[perception(rain)]
UND(undecided) ====>
[fact(rain)] [fact(rain)]
[neg, fact(rain)][speed(60)]
[speed(100)] [speed(60)]
[perm, [overtaking]]

%**** Example 2 ****
IN(accepted) =======>
[speed(60)][speed(60)]
[obl, [max_speed(100)]]
[fact(rain)][fact(rain)]
[neg, belief(agent2, rain)]
[belief(agent1, rain)]
[perception(rain)]
OUT (rejected) ======>
[speed(100)][neg, fact(rain)]
[perm, [overtaking]]

%**** Example 3 ****
IN(accepted) =======>
[speed(100)][speed(100)]
[perm, [overtaking]]
[obl, [max_speed(100)]]
[neg, fact(rain)][neg, fact(rain)]
[neg, belief(agent2, rain)]
[belief(agent1, rain)]
[neg, perception(rain)]
[perm, [overtaking]]
OUT (rejected) ======>
[speed(60)][fact(rain)]
```

Fig. 2. Example of system knowledge in the self-driving cars scenario, implemented in Arg2P (left). Arg2P system outcomes in three discussed examples (right).

The autonomous cars example, despite its simplicity, points out one of the key aspects discussed in Sect. 2: in fact, it is clear that the conversation, aimed at explaining and reaching an agreement necessarily has to involve also agent-to-agent communication with no necessarily humans in the loop. In the example, the two vehicles have an opposite perception of the surrounding environment – one thinks that it rains and one does not – so, they must converse and argue to understand what is the most possible matter of facts and then stick to that.

The examples discussed are just a simplification of the scenario but already illustrate the potential of rooting explanation in LP and argumentation. A first explanation is provided by the argumentation labelling which allows correlating arguments (and statements) accepted as plausible to a graph of attacks, superiority and non-defeasible rules, detailing the system reasoning. If we think about how the scenario could be enriched through abducibles and counterfactuals enabling a what-if analysis of different scenarios, the possibilities of the system to be explainable become manifold. Furthermore, probabilistic concepts make it possible to stick weight on assumptions, rules and arguments, for instance, agents' beliefs can be weighted according to the social credibility of each of them—possibly measured on numbers of sanctions or whatever. Ethics behaviours can be computed as well – in a human-readable way – preferring, for instance, to minimise the number of deaths in case of accidents. Interesting discussions on the moral choices of the system can be introduced and compared exploiting what-if analysis.

6 Related Works and Discussion

Just as AI sub-symbolic techniques are gaining momentum, symbolic techniques rooted in logic approaches are getting more and more attention—mostly because symbolic approaches can more easily meet the requirements of intelligent systems in terms of ethical concerns, explainability, and understandability.

Several efforts have been made for the integration of symbolic and sub-symbolic techniques under the XAI perspective, as discussed in some existing survey [9]. However, most of the works focus on a single type of logic that can effectively address the specific needs of the application at hand; less attention is typically reserved to devise a comprehensive integrated framework.

Logic programming is already undergoing a re-interpretation from an ethical perspective, as discussed in the literature [34,35,40]. Despite that, neither a comprehensive architecture nor a general methodology for integrating sub-symbolic AI techniques and logic programming can be found in the literature.

Our work aims at igniting a discussion about the integration of different logic approaches and AI techniques to achieve the XAI objectives. However, the architecture proposed in this paper is just the starting point for the design and the implementation of the envisioned symbolic engine and its integration with other existing AI techniques. Many issues and research challenges are still open.

First of all, the model formalisation deserves attention. Argumentation, *per se*, has been seen as an effective means to facilitate many aspects of decision-making and decision-support systems especially when decisions recommended by such systems need to be explained. Several works show its effectiveness [14,19], in particular in the generation of an explanation that justifies the solution found by the black boxes of ML [17,46]. Some works exist in the most recent literature, that underline the possible synergies of integration of different logic approaches, especially when combined with statistical ML algorithms. For instance, [26] discuss abduction and argumentation as two principled forms for reasoning and fleshes out the fundamental role that they can play within ML surveying the main works in the area. More generally, abduction and argumentation have been combined in different ways in the literature, starting from Dung's foundational work [18] introducing the preferred extension semantics of abductive logic programs. Abduction and argumentation can both be seen as processes for generating explanations either for a given observation as in the case of abduction or for a conclusion (claim or decision) in the case of argumentation. Explanations under abduction are in terms of underlying (theoretical or non-observable) hypotheses, whereas explanations under argumentation are in terms of arguments (among a set of known ones) that provide justified reasons for a conclusion to hold. Once again, however, a reference model does not emerge and often theoretical formalisations are not reified into any working technology.

A model for the integration of abduction and PLP is discussed in [22]; other works deal with the integration of abduction and induction [31]. In spite of the number of research activities on the subject, most approaches are scattered, and ad-hoc to solve a specific application need, and a general, well-founded framework for the integration of these models is still missing today. As a result, we are still a long way from the reification into a powerful technology – or integration of several technologies – that would allow its effective use in intelligent systems.

Moreover, for a well-founded coherent integration of all the approaches mentioned in Sect. 4, knowledge extraction and injection techniques have to be explored. A first overview of the main existing techniques is provided by [9], but some challenges remain open—in particular, knowledge injection and extraction when dealing with neural networks are a huge problem *per se*, and it is not clear *how* and *where* to inject the symbolic knowledge in nets [27].

Finally, as mentioned in Sect. 2, one of the main problems in the XAI and ethical AI field is that of selecting a satisfactory explanation. As far as the explanation selection is concerned, cognitive limitations come into play. For instance, if the explanation is required by a human, due to humans' cognitive limitations, the explanation cannot be presented with the whole chain of causal connections explaining a given algorithmic decision, rather users demand a synthetic explanation going to the core of the causal chain. Therefore, depending on the explanation's receiver the techniques to be applied may be different and could require subsequent refinement to make them cognitively understandable. The individuation of the mechanisms behind such a selection is, however, far from trivial, and many cognitive techniques should be taken into account [29].

To sum up, logic and symbolic approaches in general, especially once well integrated, can certainly be the turning point in the design of explainable and ethical systems. Nevertheless, a lot of work has to be done to ensure a well-founded integration between logic and the other AI techniques.

7 Conclusion

The paper presents a vision of how explainability and ethical behaviours in AI systems can be linked to logical concepts that find their roots in logic programming, argumentation, abduction, probabilistic LP, and inductive LP. The proposed solution is based on a (micro-)engine for injecting symbolic intelligence where and when needed. A simple example is discussed in the scenario of the self-driving car, along with its reification on a (yet preliminary) technology—namely Arg2P. However, the discussion and the corresponding example already highlight the potential benefits of the approach, once it is fruitfully integrated with the sub-symbolic models and techniques exploited in the AI field. In particular, the analysis carried out in the paper points out the key requirements of explainable and ethical autonomous behaviour, and relates them to specific logic-based approaches. The results presented here represent just a preliminary exploration of the intersection between LP and explainability: yet we think they have the potential to work as a starting point for further research.

References

1. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: results from a systematic literature review. In: 18^{th} International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2019), pp. 1078–1088. IFAAMAS, May 2019. https://dl.acm.org/doi/10.5555/3306127.3331806
2. Arrieta, A.B., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
3. Belle, V.: Symbolic logic meets machine learning: a brief survey in infinite domains. In: Davis, J., Tabia, K. (eds.) SUM 2020. LNCS (LNAI), vol. 12322, pp. 3–16. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58449-8_1
4. Borning, A., Maher, M.J., Martindale, A., Wilson, M.: Constraint hierarchies and logic programming. In: Levi, G., Martelli, M. (eds.) 6th International Conference on Logic Programming, vol. 89, pp. 149–164. MIT Press, Lisbon, Portugal (1989)
5. Calegari, R.: Micro-intelligence for the IoT: logic-based models and technologies. Ph.D. thesis, Alma Mater Studiorum-Università di Bologna, Bologna, Italy (2018). https://doi.org/10.6092/unibo/amsdottorato/8521
6. Calegari, R., Ciatto, G., Denti, E., Omicini, A.: Engineering micro-intelligence at the edge of CPCS: design guidelines. In: Montella, R., Ciaramella, A., Fortino, G., Guerrieri, A., Liotta, A. (eds.) IDCS 2019. LNCS, vol. 11874, pp. 260–270. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-34914-1_25
7. Calegari, R., Ciatto, G., Denti, E., Omicini, A.: Logic-based technologies for intelligent systems: state of the art and perspectives. Information **11**(3), 1–29 (2020). https://doi.org/10.3390/info11030167
8. Calegari, R., Ciatto, G., Mascardi, V., Omicini, A.: Logic-based technologies for multi-agent systems: a systematic literature review. Auton. Agent. Multi-Agent Syst. **35**(1), 1–67 (2020). https://doi.org/10.1007/s10458-020-09478-3
9. Calegari, R., Ciatto, G., Omicini, A.: On the integration of symbolic and sub-symbolic techniques for XAI: a survey. Intelligenza Artificiale **14**(1), 7–32 (2020). https://doi.org/10.3233/IA-190036
10. Calegari, R., Contissa, G., Lagioia, F., Omicini, A., Sartor, G.: Defeasible systems in legal reasoning: a comparative assessment. In: Araszkiewicz, M., Rodríguez-Doncel, V. (eds.) Legal Knowledge and Information Systems. JURIX 2019: The Thirty-second Annual Conference, Frontiers in Artificial Intelligence and Applications, vol. 322, pp. 169–174. IOS Press, 11–13 December 2019. https://doi.org/10.3233/FAIA190320

11. Calegari, R., Denti, E., Dovier, A., Omicini, A.: Extending logic programming with labelled variables: model and semantics. Fund. Inform. **161**(1–2), 53–74 (2018). https://doi.org/10.3233/FI-2018-1695
12. Calegari, R., Denti, E., Mariani, S., Omicini, A.: Logic programming as a service. Theory Pract. Logic Program. **18**(3–4), 1–28 (2018). https://doi.org/10.1017/S1471068418000364
13. Calegari, R., Omicini, A., Sartor, G.: Computable law as argumentation-based MAS. In: Calegari, R., Ciatto, G., Denti, E., Omicini, A., Sartor, G. (eds.) WOA 2020–21st Workshop "From Objects to Agents". CEUR Workshop Proceedings, vol. 2706, pp. 54–68. Sun SITE Central Europe, RWTH Aachen University, Aachen, Germany, October 2020. http://ceur-ws.org/Vol-2706/paper10.pdf
14. Caminada, M.: Argumentation semantics as formal discussion. J. Appl. Logics **4**(8), 2457–2492 (2017)
15. Ciatto, G., Calegari, R., Omicini, A., Calvaresi, D.: Towards XMAS: eXplainability through Multi-Agent Systems. In: Savaglio, C., Fortino, G., Ciatto, G., Omicini, A. (eds.) AI&IoT 2019 - Artificial Intelligence and Internet of Things 2019, CEUR Workshop Proceedings, vol. 2502, pp. 40–53. Sun SITE Central Europe, RWTH Aachen University, November 2019. http://ceur-ws.org/Vol-2502/paper3.pdf
16. Ciatto, G., Schumacher, M.I., Omicini, A., Calvaresi, D.: Agent-based explanations in AI: towards an abstract framework. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) Explainable, Transparent Autonomous Agents and Multi-Agent Systems. LNCS, vol. 12175, pp. 3–20. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51924-7_1
17. Cyras, K., Letsios, D., Misener, R., Toni, F.: Argumentation for explainable scheduling. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 2752–2759 (2019). https://doi.org/10.1609/aaai.v33i01.33012752
18. Dung, P.M.: Negations as hypotheses: an abductive foundation for logic programming. In: International Conference on Logic Programming, vol. 91, pp. 3–17 (1991)
19. Dung, P.M., Mancarella, P., Toni, F.: Computing ideal sceptical argumentation. Artif. Intell. **171**(10–15), 642–674 (2007). https://doi.org/10.1016/j.artint.2007.05.003
20. Dyckhoff, Roy, Herre, Heinrich, Schroeder-Heister, Peter (eds.): ELP 1996. LNCS, vol. 1050. Springer, Heidelberg (1996). https://doi.org/10.1007/3-540-60983-0
21. Esposito, F., Fanizzi, N., Iannone, L., Palmisano, I., Semeraro, G.: A counterfactual-based learning algorithm for $\mathcal{ALC}$ description logic. In: Bandini, S., Manzoni, S. (eds.) AI*IA 2005. LNCS (LNAI), vol. 3673, pp. 406–417. Springer, Heidelberg (2005). https://doi.org/10.1007/11558590_41
22. Ferilli, S.: Extending expressivity and flexibility of abductive logic programming. J. Intell. Inf. Syst. **51**(3), 647–672 (2018). https://doi.org/10.1007/s10844-018-0531-6
23. Fernández, R.R., de Diego, I.M., Aceña, V., Fernández-Isabel, A., Moguerza, J.M.: Random forest explainability using counterfactual sets. Inf. Fusion **63**, 196–207 (2020). https://doi.org/10.1016/j.inffus.2020.07.001
24. Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., Giannotti, F.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2019). https://doi.org/10.1145/3236009
25. Hulstijn, J., van der Torre, L.W.: Combining goal generation and planning in an argumentation framework. In: Hunter, A. (ed.) International Workshop on Non-monotonic Reasoning (NMR 2004), pp. 212–218. Pacific Institute, Whistler, Canada, January 2004

26. Kakas, A., Michael, L.: Abduction and argumentation for explainable machine learning: A position survey. arXiv preprint arXiv:2010.12896 (2020)
27. Kemker, R., McClure, M., Abitino, A., Hayes, T., Kanan, C.: Measuring catastrophic forgetting in neural networks. In: McIlraith, S.A., Weinberger, K.Q. (eds.) AAAI Conference on Artificial Intelligence. pp. 3390–3398. AAAI Press (2018). https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16410
28. Mariani, S., Omicini, A.: Coordination in situated systems: engineering MAS environment in TuCSoN. In: Fortino, G., Di Fatta, G., Li, W., Ochoa, S., Cuzzocrea, A., Pathan, M. (eds.) IDCS 2014. LNCS, vol. 8729, pp. 99–110. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11692-1_9
29. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
30. Modgil, S., Caminada, M.: Proof theories and algorithms for abstract argumentation frameworks. In: Simari, G., Rahwan, I. (eds.) Argumentation in artificial intelligence, pp. 105–129. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-98197-0_6
31. Mooney, R.J.: Integrating abduction and induction in machine learning. In: Flach, P.A., Kakas, A.C. (eds.) Abduction and Induction, pp. 181–191. Springer (2000). https://doi.org/10.1007/978-94-017-0606-3_12
32. Omicini, A.: Not just for humans: explanation for agent-to-agent communication. In: Vizzari, G., Palmonari, M., Orlandini, A. (eds.) AIxIA 2020 DP - AIxIA 2020 Discussion Papers Workshop. AI*IA Series, vol. 2776, pp. 1–11. Sun SITE Central Europe, RWTH Aachen University, Aachen, Germany, November 2020. http://ceur-ws.org/Vol-2776/paper-1.pdf
33. Omicini, A., Calegari, R.: Injecting (micro)intelligence in the IoT: logic-based approaches for (M)MAS. In: Lin, D., Ishida, T., Zambonelli, F., Noda, I. (eds.) MMAS 2018. LNCS (LNAI), vol. 11422, pp. 21–35. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20937-7_2
34. Pereira, L.M., Saptawijaya, A.: Programming Machine Ethics. SAPERE, vol. 26. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-29354-7
35. Pereira, L.M., Saptawijaya, A.: Counterfactuals, logic programming and agent morality. In: Urbaniak, R., Payette, G. (eds.) Applications of Formal Philosophy. LAR, vol. 14, pp. 25–53. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58507-9_3
36. Pisano, G., Calegari, R., Omicini, A., Sartor, G.: Arg-tuProlog: a tuProlog-based argumentation framework. In: Calimeri, F., Perri, S., Zumpano, E. (eds.) CILC 2020 - Italian Conference on Computational Logic. Proceedings of the 35th Italian Conference on Computational Logic. CEUR Workshop Proceedings, vol. 2719, pp. 51–66. Sun SITE Central Europe, RWTH Aachen University, CEUR-WS, Aachen, Germany, 13–15 October 2020. http://ceur-ws.org/Vol-2710/paper4.pdf
37. Poole, D.: Logic programming, abduction and probability. N. Gener. Comput. **11**(3–4), 377 (1993). https://doi.org/10.1007/BF03037184
38. Riveret, R., Oren, N., Sartor, G.: A probabilistic deontic argumentation framework. Int. J. Approximate Reasoning **126**, 249–271 (2020). https://doi.org/10.1016/j.ijar.2020.08.012
39. Rosenfeld, A., Richardson, A.: Explainability in human–agent systems. Auton. Agent. Multi-Agent Syst. **33**(6), 673–705 (2019). https://doi.org/10.1007/s10458-019-09408-y
40. Saptawijaya, A., Pereira, L.M.: From logic programming to machine ethics. In: Bendel, O. (ed.) Handbuch Maschinenethik. LAR, pp. 209–227. Springer, Wiesbaden (2019). https://doi.org/10.1007/978-3-658-17483-5_14

41. Stone, P., Veloso, M.: Multiagent systems: a survey from a machine learning perspective. Auton. Robot. **8**(3), 345–383 (2000). https://doi.org/10.1023/A:1008942012299
42. Vranes, S., Stanojevic, M.: Integrating multiple paradigms within the blackboard framework. IEEE Trans. Softw. Eng. **21**(3), 244–262 (1995). https://doi.org/10.1109/32.372151
43. Wellman, H.M.: The Child's Theory of Mind. The MIT Press (1992)
44. Wooldridge, M.J., Jennings, N.R.: Intelligent agents: theory and practice. Knowl. Eng. Rev. **10**(2), 115–152 (1995). https://doi.org/10.1017/S0269888900008122
45. Xhafa, F., Patnaik, S., Tavana, M. (eds.): IISA 2019. AISC, vol. 1084. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-34387-3
46. Zhong, Q., Fan, X., Luo, X., Toni, F.: An explainable multi-attribute decision model based on argumentation. Expert Syst. Appl. **117**, 42–61 (2019). https://doi.org/10.1016/j.eswa.2018.09.038