Towards an Explainable Argumentation-based Agent

Mariela Morveli-Espinoza and Cesar Augusto Tacla
Program in Electrical and Computer Engineering (CPGEI)
Federal University of Technology of Parana (UTFPR)
Curitiba, Brazil
morveli.espinoza@gmail.com, tacla@utfpr.edu.br

Abstract

Explainable Artificial Intelligence (XAI) systems, including intelligent agents, must be able to explain their internal decisions, behaviours, and reasoning that produce their choices to the humans (or other systems) with which they interact. In this paper, we present an initial proposal of an argumentation-based agent architecture and a mechanism for generating explanations about the goals agents commit to. We base on an extended model of BDI (Beliefs-Desires-Intentions) named the Belief-based Goal Processing (BBGP) model, which provides a more fine-grained analysis of the goal processing, which in turn may improve and enrich the informational quality of the explanations. Our proposal is based on argumentation theory, we use arguments to represent the reasons that lead an agent to make a decision and use argumentation semantics to determine acceptable arguments (reasons). Regarding the explanations, we propose two types of explanations: the partial one and the complete one. Finally, we present the future research directions with respect to both the goal processing and the explainability skills.

1 Introduction

Explainability of intelligent agents has gained attention in recent years due to their growing utilization in human applications such as recommendation or coaching systems. In these applications, the outcomes returned by the agent-based systems can be negatively affected due to the lack of explainability about their dynamics and rationality. Thus, if these systems would have explainability abilities, then their understanding, reliability, and acceptance could be enhanced.

The BDI model [Bra87][RG95] has become possibly the best-known and best-studied model of practical reasoning agents. BDI agents are able to select the goals they are going to commit to – which are called intentions – from a set of desires; however, they are not endowed with explainability abilities. In order to better understand this problem, consider a scenario of a natural disaster, where a set of robot agents wander an area in search of people needing help. When a person is seriously injured, he/she must be taken to the hospital; otherwise, he/she must be sent to a shelter. After the rescue work, the robots can be asked – by another robot or by a human – for an explanation of why a wounded person was sent to the shelter instead of taking him/her to the hospital or why the robot decided to take to the hospital a person x first, instead of taking another person

In: Rudolph S., Marreiros G. (eds.): Proceedings of STAIRS, Santiago de Compostela, Spain, august 2020, published at http://ceur-ws.org

y. In this case, it is clear that it is important to endow the agents with the ability of explain their decisions, that is, to explain how and why a certain desire became (or not) an intention.

In the case of BDI agents, the path of this explanation is made up of only one step, which happens because in BDI agents there are only two stages in the intention formation process: desires and intentions. This means that there is a lack of a fine-grained analysis of this process, which may improve and enrich the informational quality of the explanations. An extended model for intention formation has been proposed by Castelfranchi and Paglieri [CP07], which was named the Belief-based Goal Processing model (let us denote it by BBGP). The BBGP model has four stages: (i) activation (denoted ac)[1], (ii) evaluation (denoted ev), (iii) deliberation (denoted de), and (iv) checking (denoted ck). Consequently, four different statuses for a goal are defined: (i) active (=desire and denoted ac), (ii) pursuable (denoted pu), (iii) chosen (=future-directed intention and denoted ch) and (iv) executive (=present-directed intention and denoted ex). When a goal passes the activation (resp. evaluation, deliberation, checking) stage, it becomes active (resp. pursuable, chosen, executive).

Figure 1 shows a general schema of the goal processing stages, the necessary beliefs, and the status of a goal after passing each stage. In this article, a status before the active one is also considered, it is called sleeping status[2].

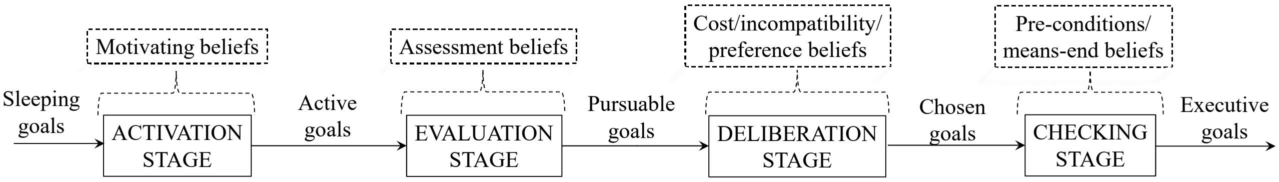

Figure 1: Schema of the goal processing stages and the status of a goal after passing each stage.

Against this background, the aim of this article is to formalize the BBGP model from the activation stage until the checking stage and to endow BBGP-based agents with explainability abilities. Thus, the research questions that are addressed in this article are: (i) how to formalize the BBGP model by integrating the activation, evaluation, deliberation, and checking stages? and (ii) how to endow BBGP-based agents with the ability of generating explanations?

In addressing the first question, we will use argumentation, which is a process of constructing and comparing arguments considering the conflicts – which are called attacks – among them. The output of argumentation process is a set (or sets) of arguments – called extensions – which are internally consistent [Dun95]. In the intention formation process or goal processing, arguments can represent reasons for a goal to change (or not) its status. Thus, one can see the intention formation process as a decision-making process, where an agent has to decide which goal(s) passes a given stage and which does not. Adopting an argumentation-based approach in a decision making problem has some benefits on the explainability. For example, a (human) user will obtain a "good" choice along with the reasons underlying this recommendation. Besides, argumentation-based decision-making is more similar with the way humans deliberate and finally make a choice [OMT10]. Regarding the second question, we endow agents with an structure to save the reasoning path, based on which explanations can be generated.

Figure 2 shows an overview of our approach, concretely it shows all the possible transitions of a goal from its sleeping status until it becomes executive, which depends on the arguments generated in each stage. We also consider the status cancelled, which happens under some circumstances that are briefly mentioned in the legend.

Next section presents the work that has been done so far, including the formalization of argumentation-based agents and our proposal for generating partial and complete explanations. Section 3 enumerates the main future direction of our research. Section 4 presents the main related work. Finally, Section 5 is devoted to conclusions.

2 What has been done so far

In this section, we present the work that has been done so far. Thus, we introduce (i) the building blocks for both the argumentation-based formalization and the mechanism for generating explanations, (ii) the argumentation process for goal processing, and (ii) the mechanism for the generation of complete and partial explanations[3].

[1] Hereafter, these notations are used for differentiate the stages and the statuses of goals.

[2] Sleeping is a status of a goal that is proposed in [Cas08] to refer to goals that have not been activated yet.

[3] The first two points were presented in [MEPPGT19] and the last point was presented in [MEPT19].

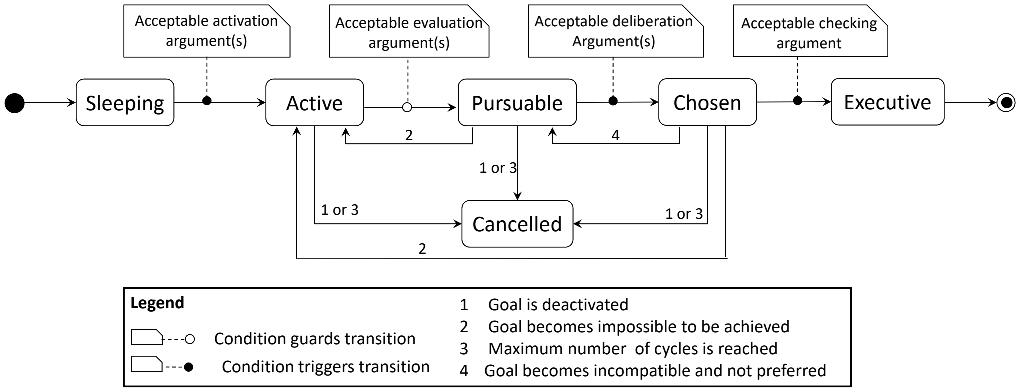
Acceptable activation argument(s)
Acceptable evaluation argument(s)
Acceptable deliberation Argument(s)
Acceptable checking argument
Sleeping
Active
Pursuable
Chosen
Executive
Cancelled
2
4
1 or 3
1 or 3
1 or 3
2
Legend
Condition guards transition
Condition triggers transition
1 Goal is deactivated
2 Goal becomes impossible to be achieved
3 Maximum number of cycles is reached
4 Goal becomes incompatible and not preferred

Figure 2: Life-cycle of goals.

2.1 Preliminaries

In this paper, BBGP-based agents use rule-based systems[4] as their basic reasoning model. The underlying logical language – denoted by $\mathcal{L}$ – consists of a set of literals[5] in first-order logical language. We represent non-ground formulae with Greek letters $(\varphi, \psi, ...)$, variables with Roman letters $(x, y, ...)$ and we name rules with $r_1, r_2, ...$. Strict rules are of the form $r = \varphi_1, ..., \varphi_n \to \psi$, and defeasible rules are of the form $r = \varphi_1, ..., \varphi_n \Rightarrow \psi$. Thus, a theory is a triple $\mathcal{T} = \langle \mathcal{F}, \mathcal{S}, \mathcal{D} \rangle$ where $\mathcal{F} \subseteq \mathcal{L}$ is a set of facts, $\mathcal{S}$ is a set of strict rules, and $\mathcal{D}$ is a set of defeasible rules. New information is produced from a given theory by applying the following concept, which was given in [AB13].

Definition 1. (Derivation schema) Let $\mathcal{T} = \langle \mathcal{F}, \mathcal{S}, \mathcal{D} \rangle$ be a theory and $\psi \in \mathcal{L}$. A derivation schema for ψ from $\mathcal{T}$ is a finite sequence $T = \{(\varphi_1, r_1), ..., (\varphi_n, r_n)\}$ such that:
- $\varphi_n = \psi$
- for $i = 1...n$, $\varphi_i \in \mathcal{F}$ and $r_i = \emptyset$, or $r_i \in \mathcal{S} \cup \mathcal{D}$

Based on a derivation scheme T, the following sets can be defined: $\texttt{SEQ}(T) = \{\varphi_1, ..., \varphi_n\}$, $\texttt{FACTS}(T) = \{\varphi_i | i \in \{1, ..., n\}, r_i = \emptyset\}$, $\texttt{STRICT}(T) = \{r_i | i \in \{1, ..., n\}, r_i \in \mathcal{S}\}$, $\texttt{DEFE}(T) = \{r_i | i \in \{1, ..., n\}, r_i \in \mathcal{D}\}$

2.2 Building Blocks

From $\mathcal{L}$, we distinguish the following finite sets:
- $\mathcal{F}$ is the set of facts of the agent and
- $\mathcal{G}$ is the set of goals of the agent.

$\mathcal{F}$ and $\mathcal{G}$ are subsets of ground literals from the language $\mathcal{L}$ and are pairwise disjoint. Besides, $\mathcal{G} = \mathcal{G}_{\mathsf{ac}} \cup \mathcal{G}_{\mathsf{pu}} \cup \mathcal{G}_{\mathsf{ch}} \cup \mathcal{G}_{\mathsf{ex}}$, where $\mathcal{G}_{\mathsf{ac}}$ (resp. $\mathcal{G}_{\mathsf{pu}}$, $\mathcal{G}_{\mathsf{ch}}$, $\mathcal{G}_{\mathsf{ex}}$) stands for the set of active (resp. pursuable, chosen, executive) goals. It holds that $\mathcal{G}_x \cap \mathcal{G}_y = \emptyset$, for $x, y \in \{\mathsf{ac}, \mathsf{pu}, \mathsf{ch}, \mathsf{ex}\}$ with $x \neq y$.

Other important structures are rules, which express the relation between beliefs and goals. Rules can be classified in standard and non-standard rules (activation, evaluation, deliberation, and checking rules). The former are made up of beliefs in both their premises and their conclusions and the latter are made up of beliefs in their premises and goals or beliefs about goals in their conclusions. Both standard and non-standard rules can be strict or defeasible rules. Standard rules can be used in any of the stages of the goal processing whereas non-standard rules are distinct for each stage. Thus, we have:

- ***Standard rules*** (r_{st}): $\bigwedge \varphi_i \to \phi$ (or $\bigwedge \varphi_i \Rightarrow \phi$).
- ***Activation rules*** (r_{ac}): $\bigwedge \varphi_i \to \psi$ (or $\bigwedge \varphi_i \Rightarrow \psi$).
- ***Evaluation rules*** (r_{ev}): $\bigwedge \varphi_i \to \neg\psi$ (or $\bigwedge \varphi_i \Rightarrow \neg\psi$).
- ***Deliberation rules***: $r^1_{de} = \neg has\_incompatibility(g) \to chosen(g)$
  $r^2_{de} : most\_valuable(g) \to chosen(g)$
- ***Checking rule***: $r_{ck} = has\_plans\_for(g) \wedge satisfied\_context\_for(g) \to executive(g)$

Where φ_i and ψ denote non-ground literals that represent beliefs and goals, respectively. g denote ground

[4] Rule-based systems distinguish between facts, strict rules, and defeasible rules. A strict rule encodes strict information that has no exception, whereas a defeasible rule expresses general information that may have exceptions.

[5] Literals are defined as positive or negative atoms where an atom is an n-ary predicate.

literals that represent a goal[6]. Notice that standard, activation, and evaluation rules are designed and entered by the programmer of the agent, and their content is domain-dependent. Otherwise, rules in deliberation and checking stages are pre-defined and domain-independent. Finally, let $\mathcal{R}_{st}, \mathcal{R}_{ac}, \mathcal{R}_{ev}, \mathcal{R}_{de}$, and $\mathcal{R}_{ck}$ denote the set of standard, activation, evaluation, deliberation, and checking rules, respectively. Next, we define the theory of a BBGP-based agent.

Definition 2. (BBGP-based Agent Theory) A theory is a triple $\mathcal{T} = \langle \mathcal{F}, \mathcal{S}, \mathcal{D} \rangle$ such that: (i) $\mathcal{F}$ is the set of beliefs of the agent, (ii) $\mathcal{S} = \mathcal{R}^{\mathcal{S}}_{st} \cup \mathcal{R}^{\mathcal{S}}_{ac} \cup \mathcal{R}^{\mathcal{S}}_{ev} \cup \mathcal{R}^{\mathcal{S}}_{de} \cup \mathcal{R}^{\mathcal{S}}_{ck}$ is the set of strict rules, and (ii) $\mathcal{D} = \mathcal{R}^{\mathcal{D}}_{st} \cup \mathcal{R}^{\mathcal{D}}_{ac} \cup \mathcal{R}^{\mathcal{D}}_{ev} \cup \mathcal{R}^{\mathcal{D}}_{de} \cup \mathcal{R}^{\mathcal{D}}_{ck}$ is the set of defeasible rules, where $\mathcal{R}_x = \mathcal{R}^{\mathcal{S}}_x \cup \mathcal{R}^{\mathcal{D}}_x$ (for $x \in \{st, ac, ev, de, ck\}$). It holds that $\mathcal{R}^{\mathcal{S}}_x \cap \mathcal{R}^{\mathcal{D}}_x = \emptyset$.

From a theory, a BBGP-based agent can build arguments. There are two categories of arguments. The first one – called epistemic arguments – justifies or attacks beliefs, while the other one – called stage arguments – justifies or attacks the passage of a goal from one stage to another. There is a set of arguments for each stage of the BBGP model.

Definition 3. (Arguments) Let $\mathcal{T} = \langle \mathcal{F}, \mathcal{S}, \mathcal{D} \rangle$ be a BBGP-based agent theory and $\mathcal{T}' = \langle \mathcal{F}, \mathcal{R}^{\mathcal{S}}_{st}, \mathcal{R}^{\mathcal{D}}_{st} \rangle$ and $\mathcal{T}'' = \langle \mathcal{F}, \mathcal{S}'', \mathcal{D}'' \rangle$ be two sub-theories of $\mathcal{T}$, where $\mathcal{S}' = \mathcal{S} \setminus \mathcal{R}^{\mathcal{S}}_{st}$ and $\mathcal{D}' = \mathcal{D} \setminus \mathcal{R}^{\mathcal{D}}_{st}$. An ***epistemic argument*** constructed from $\mathcal{T}'$ is a pair $A = \langle T, \varphi \rangle$ such that:
(1) $\varphi \in \mathcal{L}$
(2) T is a derivation schema for φ from $\mathcal{T}'$
On the other hand, a ***stage argument*** constructed from $\mathcal{T}''$ is a pair $A = \langle T, g \rangle$ such that:
(1) $g \in \mathcal{G}$
(2) For the activation and evaluation stages: T is a derivation schema for g from $\mathcal{T}''$. For the deliberation stage: T is a derivation schema for $chosen(g)$ from $\mathcal{T}''$. For the checking stage: T is a derivation schema for $executive(g)$ from $\mathcal{T}''$.

For both kinds of arguments, it holds that $\texttt{SEQ}(T)$ is consistent[7] and T must be minimal[8]. Finally, $\texttt{ARG}_{ep}$, $\texttt{ARG}_{ac}$, $\texttt{ARG}_{ev}$, $\texttt{ARG}_{de}$, and $\texttt{ARG}_{ck}$ denote the set of all epistemic, activation, evaluation, deliberation, and checking arguments, respectively. As for notation, $\texttt{CLAIM}(A) = \varphi$ (or g) and $\texttt{SUPPORT}(A) = T$ denote the conclusion and the support of an argument A, respectively.

An argument may have a set of sub-arguments. Thus, an argument $\langle T', \varphi' \rangle$ is a ***sub-argument*** of $\langle T, \varphi \rangle$ iff $\texttt{FACTS}(T') \subseteq \texttt{FACTS}(T), \texttt{STRICT}(T') \subseteq \texttt{STRICT}(T)$, and $\texttt{DEFE}(T') \subseteq \texttt{DEFE}(T)$. $\texttt{SUB}(A)$ denotes the set of all sub-arguments of A.

Stage arguments built from $\mathcal{T}$ constitute a cause for a goal changes its status. However, it is not a proof that the the goal should adopt another status. The reason is that an argument can be attacked by other arguments. Two kinds of attacks are distinguished: (i) the attacks between epistemic arguments, and (ii) the mixed attacks, in which an epistemic argument attacks a stage argument. The former is defined over $\texttt{ARG}_{ep}$ and is captured by the binary relation $\mathsf{att}_{ep} \subseteq \texttt{ARG}_{ep} \times \texttt{ARG}_{ep}$. The latter is defined over $\texttt{ARG}_{ep}$ and $\texttt{ARG}_x$ (for $x \in \{ac, ev, de, ck\}$); and is captured by the binary relation $\mathsf{att}_{mx} \subseteq \texttt{ARG}_{ep} \times \texttt{ARG}_x$. For both kinds of attacks, $(A, B) \in \mathsf{att}_{mx}$ (or $(A, B) \in \mathsf{att}_{ep}$) denotes that there is an attack relation between arguments A and B. Next definition captures both kinds of attacks; thus, rebuttal may occur only between epistemic arguments and undercut may occur in both kinds of attacks.

Definition 4. (Attacks) Let $\langle T', \varphi' \rangle$ and $\langle T, \varphi \rangle$ be two epistemic arguments, and $\langle T, g \rangle$ be a stage argument. $\langle T', \varphi' \rangle$ ***rebuts*** $\langle T, \varphi \rangle$ if $\varphi = \neg \varphi'$. $\langle T, \varphi \rangle$ ***undercuts*** $\langle T', \neg \varphi' \rangle$ (or $\langle T, g \rangle$) if $\varphi' \in \texttt{FACTS}(T)$.

From epistemic and stage arguments and the attacks between them, it is generated a different Argumentation Framework (AF) for each stage of the BBGP model.

Definition 5. (Argumentation Framework) An argumentation framework $\mathcal{AF}_x$ is a pair $\mathcal{AF}_x = \langle \texttt{ARG}, \mathsf{att} \rangle$ ($x \in \{ac, ev, de, ck\}$) such that:

- $\texttt{ARG} = \texttt{ARG}_x \cup \texttt{ARG}'_{ep} \cup \texttt{SUBARGS}$, where $\texttt{ARG}_x$ is a set of stage arguments, $\texttt{ARG}'_{ep} = \{A \mid A \in \texttt{ARG}_{ep}$ and $(A, B) \in \mathsf{att}_{mx}$ or $(A, C) \in \mathsf{att}_{ep}\}$, where $B \in \texttt{ARG}_x$ and $C \in \texttt{ARG}'_{ep}$, and $\texttt{SUBARGS} = \bigcup_{A \in \texttt{ARG}_x, A \in \texttt{ARG}'_{ep}} \texttt{SUB}(A)$.

[6] In any of the states of the goal processing, a goal is represented by a ground atom. However, before a goal becomes active, it has the form of a non-ground atom; in this case, we call it a sleeping goal. Thus, ψ is a sleeping goal and g a goal in some status.

[7] A set $\mathcal{L}' \subseteq \mathcal{L}$ is consistent iff $\nexists \varphi, \varphi' \in \mathcal{L}'$ such that $\varphi = \neg \varphi'$. It is inconsistent otherwise.

[8] Minimal means that there is no $T' \subset T$ such that φ (g, $chosen(g)$, or $executive(g)$) is a derivation schema of T'.

- $\mathsf{att} = \mathsf{att}'_{ep} \cup \mathsf{att}'_{mx}$, where $\mathsf{att}'_{ep} \subseteq \mathtt{ARG}'_{ep} \times \mathtt{ARG}'_{ep}$ and $\mathsf{att}'_{mx} \subseteq \mathtt{ARG}'_{ep} \times \mathtt{ARG}_{x}$.

The next step is to evaluate the arguments that make part of the AF. This evaluation is important because it determines which goals pass (acceptable goals) from one stage to the next. The aim is to obtain a subset of $\mathtt{ARG}$ without conflicting arguments. In order to obtain it, we use an *acceptability semantics*, which return one or more sets – called extensions – of acceptable arguments. The fact that a stage argument belong to an extension determines the change of status of the goal in its claim. Next, the main semantics introduced by Dung [Dun95] are recalled[9].

Definition 6. (Semantics) Let $\mathcal{AF}_x = \langle \mathtt{ARG}, \mathsf{att} \rangle$ be an AF (with $x \in \{ac, ev, de, ck\}$) and $\mathcal{E} \subseteq \mathtt{ARG}$:
- $\mathcal{E}$ is ***conflict-free*** if $\forall A, B \in \mathcal{E}$, $(A, B) \notin \mathsf{att}$ or $(B, A) \notin \mathsf{att}$
- $\mathcal{E}$ ***defends*** A iff $\forall B \in \mathtt{ARG}$, if $(B, A) \in \mathsf{att}$, then $\exists C \in \mathcal{E}$ s.t. $(C, B) \in \mathsf{att}$.
- $\mathcal{E}$ is ***admissible*** iff it is conflict-free and defends all its elements.
- A conflict-free $\mathcal{E}$ is a ***complete extension*** iff we have $\mathcal{E} = \{A | \mathcal{E} \text{ defends } A\}$.
- $\mathcal{E}$ is a ***preferred extension*** iff it is a maximal (w.r.t the set inclusion) complete extension.
- $\mathcal{E}$ is a ***grounded extension*** iff is a minimal (w.r.t. set inclusion) complete extension.
- $\mathcal{E}$ is a ***stable extension*** iff $\mathcal{E}$ is conflict-free and $\forall A \in \mathtt{ARG}$, $\exists B \in \mathcal{E}$ such that $(B, A) \in \mathsf{att}$.

2.3 Partial and Complete Explanations

In order to able to generate partial and complete explanations, a BBGP-based agent needs to store information about the progress of his goals, that is, the changes of the statuses of such goals and the causes of these changes. The latter are stored in each AF in form of accepted arguments; however, the former cannot be stored in an AF. Thus, we need a structure that saves both the status of each goal and the AF that supports this status. Considering that at each stage, the agent generates arguments and attacks for more than one goal – which are stored in each $\mathcal{AF}_x$ – and we only need those arguments and attacks related to one goal, we have to extract from $\mathcal{AF}_x$ such arguments and attacks. In other words, we need to obtain a sub-AF.

Definition 7. (Sub-AF) Let $\mathcal{AF}_x = \langle \mathtt{ARG}, \mathsf{att} \rangle$ (with $x \in \{ac, ev, de, ck\}$) be the an AF and $g \in \mathcal{G}$ a goal. An AF $\mathcal{AF}'_x = \langle \mathtt{ARG}', \mathsf{att}' \rangle$ is a *sub-AF* of $\mathcal{AF}_x$ (denoted $\mathcal{AF}'_x \sqsubseteq \mathcal{AF}_x$), if $\mathtt{ARG}' \subseteq \mathtt{ARG}$ and $\mathsf{att}' = \mathsf{att} \otimes \mathtt{ARG}'$, s.t.:
- $\mathtt{ARG}' = \{\{A | A \in \mathtt{ARG}_x, \mathtt{CLAIM}(A) = g\} \cup \{B | (B, A) \in \mathsf{att}_{mx} \text{ or } (B, C) \in \mathsf{att}'_{ep}\}$, where $B \in \mathtt{ARG}_{ep}$, $C \in \mathtt{ARG}'_{ep}, \mathsf{att}'_{ep} \subset \mathsf{att}'$, and $\mathtt{ARG}'_{ep} \subset \mathtt{ARG}'\}\}$, and
- $\mathsf{att} \otimes \mathtt{ARG}'$ returns a subset of att that involves just the arguments in $\mathtt{ARG}'$.

Next, we define a structure that stores the causes of the change of the status of a goal, which must be updated after a new change occurs. We can see this structure as a table record, where each row saves the status of a goal along with the AF that supports such status.

Definition 8. (Goal Memory) Let $\mathcal{AF}_x = \langle \mathtt{ARG}, \mathsf{att} \rangle$ be an AF (with $x \in \{ac, ev, de, ck\}$), $\mathcal{AF}'_x \sqsubseteq \mathcal{AF}_x$ a sub-AF, and $g \in \mathcal{G}$ a goal. The goal memory $\mathcal{GM}_g$ for goal g is a set of ordered pairs $(\mathtt{STA}, \mathtt{REASON})$ such that:

- $\mathtt{STA} \in \{\mathsf{ac}, \mathsf{pu}, \mathsf{ch}, \mathsf{ex}, \mathsf{not\ ac}, \mathsf{not\ pu}, \mathsf{not\ ch}, \mathsf{not\ ex}\}$ where $\{\mathsf{ac}, \mathsf{pu}, \mathsf{ch}, \mathsf{ex}\}$ represent the status g attains due to the arguments in $\mathtt{REASON}$ whereas $\{\mathsf{not\ ac}, \mathsf{not\ pu}, \mathsf{not\ ch}, \mathsf{not\ ex}\}$ represent the status g cannot attain due to the arguments in $\mathtt{REASON}$.
- $\mathtt{REASON} = \mathcal{AF}'_x \sqsubseteq \mathcal{AF}_x$ is a sub-AF whose selected extension supports the current status of g.

Let $\mathcal{GM}^+$ be the set of all goal memories and $\mathtt{NUM\_REC} : \mathcal{GM}^+ \to \mathbb{N}$ a function that returns the number of records of a given $\mathcal{GM}$. From the goal memory structure, the partial and complete explanation can be generated.

Definition 9. (Partial and Complete Explanations) Let $g \in \mathcal{G}$ be a goal, $\mathcal{GM}_g$ the memory of g, and $\mathcal{AF}_{ac}$, $\mathcal{AF}_{ev}$, $\mathcal{AF}_{de}$, and $\mathcal{AF}_{ck}$ the four argumentation frameworks involved in the goal processing. Besides, let $x \in \{\mathsf{ac}, \mathsf{pu}, \mathsf{ch}, \mathsf{ex}\}$ denote the current status of g:
- A **complete explanation** $\mathcal{CE}_g$ for $g \in \mathcal{G}_x$ is obtained as follows: $\mathcal{CE}_g = \bigcup_{i=1}^{i=\mathtt{NUM\_REC}(\mathcal{GM}_g)} \mathtt{REASON}_i$, where $\mathtt{REASON}_i \sqsubseteq \mathcal{AF}_{ac}$, $\mathtt{REASON}_i \sqsubseteq \mathcal{AF}_{ev}$, $\mathtt{REASON}_i \sqsubseteq \mathcal{AF}_{de}$, and $\mathtt{REASON}_i \sqsubseteq \mathcal{AF}_{ck}$.
- A **partial explanation** $\mathcal{PE}_g$ is obtained as follows: $\mathcal{PE}_g = \bigcup_{i=1}^{i=\mathtt{NUM\_REC}(\mathcal{GM}_g)} \mathcal{E}_i$, where $\mathcal{E}_i$ is the selected extension obtained from $\mathtt{REASON}_i$.

[9] It is not the scope of this article to study the most adequate semantics for goal processing or the way to select an extension when more than one is returned by a semantics.

This means that complete explanations include the arguments that support and attack the pass of a goal to the next stage whereas partial explanations only include the supporting arguments.

Example 1. Considering the scenario presented in the Introduction section, suppose that BOB is a rescue robot whose goal memory for goal $\mathsf{g}_2 = take\_hospital(patient_1)$ (that is, $\mathcal{GM}_{g_2}$) is shown in the following table[10].

STA	REASON
ac	$\mathcal{AF}^{g_2}_{ac} = \langle\{\mathsf{A}^2_{\mathsf{ac}}, \mathsf{A}^3_{\mathsf{ac}}, \mathsf{A}^7_{\mathsf{ep}}, \mathsf{A}^2_{\mathsf{ep}}, \mathsf{A}^3_{\mathsf{ep}}, \mathsf{A}^9_{\mathsf{ep}}, \mathsf{A}^5_{\mathsf{ep}}, \mathsf{A}^8_{\mathsf{ep}}\},\ \{(\mathsf{A}^8_{\mathsf{ep}}, \mathsf{A}^2_{\mathsf{ac}}), (\mathsf{A}^7_{\mathsf{ep}}, \mathsf{A}^8_{\mathsf{ep}}), (\mathsf{A}^8_{\mathsf{ep}}, \mathsf{A}^7_{\mathsf{ep}}), (\mathsf{A}^9_{\mathsf{ep}}, \mathsf{A}^8_{\mathsf{ep}})\}\rangle$
pu	$\mathcal{AF}^{g_2}_{ev} = \langle\{\mathsf{A}^1_{\mathsf{ev}}, \mathsf{A}^{11}_{\mathsf{ep}}, \mathsf{A}^6_{\mathsf{ep}}\}, \{(\mathsf{A}^{11}_{\mathsf{ep}}, \mathsf{A}^1_{\mathsf{ev}}), (\mathsf{A}^{11}_{\mathsf{ep}}, \mathsf{A}^6_{\mathsf{ep}}), (\mathsf{A}^6_{\mathsf{ep}}, \mathsf{A}^{11}_{\mathsf{ep}})\}\rangle$
ch	$\mathcal{AF}^{g_2}_{de} = \langle\{\mathsf{A}^1_{\mathsf{de}}, \mathsf{A}^{12}_{\mathsf{ep}}\}, \{\}\rangle$
ex	$\mathcal{AF}^{g_2}_{ck} = \langle\{\mathsf{A}^1_{\mathsf{ck}}, \mathsf{A}^{13}_{\mathsf{ep}}, \mathsf{A}^{11}_{\mathsf{ep}}\}, \{\}\rangle$

Now suppose that at end of a rescue day, BOB is interrogated by a supervisor with the following question: $\mathtt{WHY}(g_2, ac)$, which in natural language would be: *Why did you decide to activate the goal "taking patient$_1$ to the hospital?"*. Considering $\mathcal{GM}_{g_2}$, the complete explanation is:

$\mathcal{CE}_{g_2} = \mathcal{AF}^{g_2}_{ac} = \langle\{\mathsf{A}^2_{\mathsf{ep}}, \mathsf{A}^3_{\mathsf{ep}}, \mathsf{A}^5_{\mathsf{ep}}, \mathsf{A}^7_{\mathsf{ep}}, \mathsf{A}^8_{\mathsf{ep}}, \mathsf{A}^9_{\mathsf{ep}}, \mathsf{A}^2_{\mathsf{ac}}, \mathsf{A}^3_{\mathsf{ac}}\}, (\mathsf{A}^8_{\mathsf{ep}}, \mathsf{A}^2_{\mathsf{ac}}),\ \{(\mathsf{A}^7_{\mathsf{ep}}, \mathsf{A}^8_{\mathsf{ep}}), (\mathsf{A}^8_{\mathsf{ep}}, \mathsf{A}^7_{\mathsf{ep}}), (\mathsf{A}^9_{\mathsf{ep}}, \mathsf{A}^8_{\mathsf{ep}})\}\rangle$

In natural language, this would be the answer: $\mathsf{patient}_1$ *had a fractured bone ($\mathsf{A}^2_{\mathsf{ep}}$), the fractured bone was of his arm ($\mathsf{A}^3_{\mathsf{ep}}$), and it was an open fracture ($\mathsf{A}^5_{\mathsf{ep}}$). Given that he had a fractured bone, he might be considered severe injured ($\mathsf{A}^7_{\mathsf{ep}}$); however, since such fracture was of his arm, it might not be considered a severe injure ($\mathsf{A}^8_{\mathsf{ep}}$). Finally, I noted that it was an open fracture, which determines – without exception – that it was a severe injury ($\mathsf{A}^9_{\mathsf{ep}}$). For these reasons I decided to activate the goal taking him to the hospital ($\mathsf{A}^2_{\mathsf{ac}}, \mathsf{A}^3_{\mathsf{ac}}$).*

On the other hand, for generating the partial explanation BOB uses the arguments that are part of the a preferred extension. Thus, he answers with: $\mathcal{PE}_{g_2} = \{\mathsf{A}^2_{\mathsf{ep}}, \mathsf{A}^3_{\mathsf{ep}}, \mathsf{A}^5_{\mathsf{ep}}, \mathsf{A}^7_{\mathsf{ep}}, \mathsf{A}^9_{\mathsf{ep}}, \mathsf{A}^2_{\mathsf{ac}}, \mathsf{A}^3_{\mathsf{ac}}\}$.

In natural language he would give the following answer: $\mathsf{patient}_1$ *had a fractured bone ($\mathsf{A}^2_{\mathsf{ep}}$), the fractured bone was of his arm (A^3_{ep}), and it was an open fracture ($\mathsf{A}^5_{\mathsf{ep}}$); therefore, he was severely injured ($\mathsf{A}^7_{\mathsf{ep}}, \mathsf{A}^9_{\mathsf{ep}}$). Since he was severely injured I took him to the hospital ($\mathsf{A}^2_{\mathsf{ac}}, \mathsf{A}^3_{\mathsf{ac}}$).*

3 What still needs to be done

Next, we present some future research directions of this work:

- Considering Figure 2, goals can go forward in the intention formation process and also can go back, they can even become cancelled. For example, a chosen goal can return to a previous status like pursuable or active. We have briefly given an idea about the reasons for a goal lose its current status; however, there is no formalization about it and there is not a mechanism that controls the life-cycle depicted in Figure 2.

- We have considered an additional status (cancelled); however, it is important study if there is the need of another statuses (like suspended of [BPML04]). The inclusion of new statuses impacts on explainability skills of the agents; however, it also makes the model more complex. So, there is also a need for a further study about it and about how the dynamic between the statuses could be.

- We have focused on achievement (or procedural) goals[11]; however, maintenance and declarative goals[12] are also considered in the evaluation stage of the BBGP model. According to the BBGP model, maintenance goals are also evaluated to determine if a goal becomes or not pursuable. For example, one goal of human beings is to be happy and it is a goal that is permanently active. If a person is already happy, this goal does not becomes pursuable because it is not necessary to do something to achieve it; on the other hand, if the person is not happy, this goal becomes pursuable because the person has to do something to achieve his/her happiness. BBGP-based agents also evaluate goals that depend on other agents, that is, the achievement of an state of the world depends on the actions execution of other agents. Therefore, declarative goals are used to evaluate a certain state of the world.

[10] Due to the lack of space, we are not going to explain the entire example; however, the reader can find it in [MEPT19].

[11] A goal is called procedural when there is a set of plans for achieving it. This differs from declarative ones, which are a description of the state sought [WPHT02].

[12] Unlike achievement goals, maintenance goals define states that must remain true, rather than a state that is to be achieved [DHT06].

- During deliberation stage, BBGP-based agents have to identify possible incompatibilities between pursuable goals and resolve such incompatibilities in order to determine those goals become chosen. This problem was briefly tackled in [MEPPGT17] and extended in [MENP+19]; however, it was not integrated in the BBGP-based agents architecture. Besides, based on [MENP+19], it is possible to improve and add new deliberations rules. This in turn may enrich the explanations generated by BBGP-based agents.

- In [MENPT19], we have took into account uncertainty for identifying and dealing with incompatibilities among goals. It would be interesting to also consider uncertainty in the elements of activation, evaluation, deliberation rules rules. How would this impact on the results? How would this impact on the explainability skills of agents?

- Regarding the generated explanations, there is still a lot to work to do. We have proposed two kinds of explanations; however, it is necessary to study how to deal with complex questions, which require more elaborate and adequate explanations. In this sense, a "good" explanation may include elements from different AFs, which elements? how to organize them? What else should be taken into account for generating an explanation?

4 Related Work

Since XAI is a recently emerged domain in Artificial Intelligence, there are few reviews about the works in this area. In [ANCF19], Anjomshoae et al. make a Systematic Literature Review about goal-driven XAI, i.e., explainable agency for robots and agents. According to them, there are very few works tackle inter-agent explainability. One interesting research question was about the platforms and architectures that have been used to design Explainable Agency. Their results show that 9% implemented their explanations in BDI architecture. In [BHH+10] and [HvdBM10], the authors focus on generating explanations for humans about their goals and actions. They construct a tree with beliefs, goals, and actions, from which they explanations are constructed. Unlike our proposal, their explanations do not detail the progress of the goals and are not complete in the sense that do not express why an agent did not commit to a given goal.

Finally, Sassoon et al. [SSKP19] propose an approach of explainable argumentation based on argumentation schemes and argumentation-based dialogues. In this approach, an agent provides explanations to patients (human users) about their treatments. In this case, argumentation is applied in a different way than in our proposal and with other focus, they generate explanations for information seeking and persuasion.

5 Final Remarks

This article presented an approach for explainable agency based on argumentation theory. The chosen architecture was the BBGP model, which can be considered an extension of the BDI model. Our objective was that BBGP-based agents be able to explain their decision about the statuses of their goals, especially those goals they committed to. In order to achieve our objectives, we equipped BBGP-based agents with a structure and a mechanism to generate both partial and complete explanations.

Furthermore, we presented an agenda about some future research directions. Currently, we are working on integrating the identification and resolution of incompatibilities to the whole BBGP-based agent architecture. We are also extending the set of deliberation rules in order to produce more detailed explanations. Finally, we are working on the implementation of the initial proposal.

Acknowledgements

This work is fully founded by CAPES.

References

[AB13] Leila Amgoud and Philippe Besnard. A formal characterization of the outcomes of rule-based argumentation systems. In *International Conference on Scalable Uncertainty Management*, pages 78–91. Springer, 2013.

[ANCF19] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1078–1088, 2019.

[BHH^{+}10] Joost Broekens, Maaike Harbers, Koen Hindriks, Karel Van Den Bosch, Catholijn Jonker, and John-Jules Meyer. Do you get it? user-evaluated explainable bdi agents. In *German Conference on Multiagent System Technologies*, pages 28–39. Springer, 2010.

[BPML04] Lars Braubach, Alexander Pokahr, Daniel Moldt, and Winfried Lamersdorf. Goal representation for bdi agent systems. In *International Workshop on Programming Multi-Agent Systems*, pages 44–65. Springer, 2004.

[Bra87] Michael Bratman. Intention, plans, and practical reason. 1987.

[Cas08] Cristiano Castelfranchi. Reasons: Belief support and goal dynamics. *Mathware & Soft Computing*, 3(1-2):233–247, 2008.

[CP07] Cristiano Castelfranchi and Fabio Paglieri. The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese*, 155(2):237–263, 2007.

[DHT06] Simon Duff, James Harland, and John Thangarajah. On proactivity and maintenance goals. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 1033–1040, 2006.

[Dun95] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[HvdBM10] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. Design and evaluation of explainable bdi agents. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 125–132. IEEE, 2010.

[MENP^{+}19] Mariela Morveli-Espinoza, Juan Carlos Nieves, Ayslan Possebom, Josep Puyol-Gruart, and Cesar Augusto Tacla. An argumentation-based approach for identifying and dealing with incompatibilities among procedural goals. *International Journal of Approximate Reasoning*, 105:1–26, 2019.

[MENPT19] Mariela Morveli Espinoza, Juan Carlos Nieves, Ayslan Possebom, and Cesar Augusto Tacla. Dealing with incompatibilities among procedural goals under uncertainty. *Inteligencia Artificial. Ibero-American Journal of Artificial Intelligence*, 22(64), 2019.

[MEPPGT17] Mariela Morveli Espinoza, Ayslan T Possebom, Josep Puyol-Gruart, and Cesar A Tacla. Dealing with incompatibilities among goals. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1649–1651. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[MEPPGT19] Mariela Morveli-Espinoza, Ayslan Trevizan Possebom, Josep Puyol-Gruart, and Cesar Augusto Tacla. Argumentation-based intention formation process. *Revista Internacional Dyna*, 86(208):82–91, 2019.

[MEPT19] Mariela Morveli-Espinoza, Ayslan Possebom, and Cesar Augusto Tacla. Argumentation-based agents that explain their decisions. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 467–472. IEEE, 2019.

[OMT10] Wassila Ouerdane, Nicolas Maudet, and Alexis Tsoukias. Argumentation theory and decision aiding. In *Trends in Multiple Criteria Decision Analysis*, pages 177–208. Springer, 2010.

[RG95] Anand S Rao and Michael P Georgeff. BDI agents: from theory to practice. In *ICMAS*, volume 95, pages 312–319, 1995.

[SSKP19] Isabel Sassoon, Elizabeth Sklar, Nadin Kokciyan, and Simon Parsons. Explainable argumentation for wellness consultation. In *Proceedings of 1st International Workshop on eXplanable TRansparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS2019), AAMAS*, 2019.

[WPHT02] Michael Winikoff, Lin Padgham, James Harland, and John Thangarajah. Declarative and procedural goals in intelligent agent systems. In *International Conference on Principles of Knowledge Representation and Reasoning.* Morgan Kaufman, 2002.