Goal Reasoning and Explanations Generation in BDI-extended Agents

Mariela Morveli-Espinoza[1] , **Juan Carlos Nieves**[2] , **Cesar A. Tacla**[1] , **Henrique M. Jasinski**[1]
[1]CPGEI- Federal University of Technology of Parana (UTFPR), Curitiba - Brazil
[2]Department of Computing Science, Umeå University, Umeå - Sweden
morveli.espinoza@gmail.com, jcnieves@cs.umu.se, tacla@utfpr.edu.br,
henrique.r.monteiro@hotmail.com

Abstract

Explainable Artificial Intelligence systems, including intelligent agents, are expected to explain their internal decisions, behaviors, and reasoning that produce their choices to the humans (or to other systems) with which they interact. This work (i) formalizes a practical reasoning agent model, which is a more granular and refined than the BDI (beliefs-desires-intentions) model and (ii) endows agents with explicability abilities, whose informational quality is rich due the fine-grained details. To the best of our knowledge, it is one of the few works to equip a BDI agent with a structure and a mechanism to generate explanations.

1 Introduction

Explicability is one of the necessary ethical principles that must be respected in order to reach the trustworthiness of AI[1] systems (Smuha 2019). In intelligent agents, it has gained attention in recent years due to their growing utilization in human-AI interaction applications such as recommendation or coaching systems in domains such as e-health (e.g., (Guerrero, Nieves, and Lindgren 2016)), UAVs (Unmanned Aerial Vehicle) (e.g., (Gunetti, Thompson, and Dodd 2013)), or smart environments (e.g., (Nieves and Lindgren 2014)). In these applications, the outcomes returned by the agent-based systems can be negatively affected due to the lack of clarity and explicability about their dynamics and rationality. Thus, if these systems would be equipped with explicability abilities, then their understanding, reliability, and acceptance could be enhanced.

The BDI model (Bratman 1987) is possibly the best-known and best-studied model of practical reasoning agents. In this model, agents deliberate which actions to perform in order to achieve their goals, which are selected during the goal selection process. 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 explicability abilities. So they cannot justify how an intention was formed or describe the reasoning path that allows a desire become an intention. Explaining goal reasoning is a critical feature of rational agents because goals guide their actions and this procedural aspect of goals is essential for the practicality of agents in highly dynamic environments (Winikoff et al. 2002).

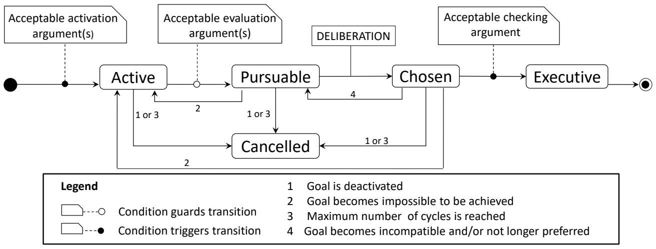
An extended model for goal processing has been proposed in (Castelfranchi and Paglieri 2007). This is the Belief-based Goal Processing Model (we will denote it by BBGP), which is a four-stage goal processing model, where the stages are: (i) activation, (ii) evaluation, (iii) deliberation, and (iv) checking. This fine-grained detail of the goal processing may have relevant consequences for the analysis of what an intention is and may better explain how an intention becomes what it is. This model makes explicit the function of beliefs in the goal processing as a diachronic support (i.e. it happens in time) and synchronic support (it leaves traces, which means that there is a memory of the cognitive path that conduced to the outcome). According to (Smuha 2019), explicability involves traceability, auditability, and transparency. The BBGP model is ideal for supporting these characteristics, the synchronic support gives traceability and auditability and the diachronic support enforces auditability because the agent is able to give the reasons – in form of beliefs – for a goal change its status.

2 Contributions

The first contribution of this work is the formalization of the goal reasoning in the BBGP-based model, which was done by using formal argumentation reasoning; in concrete, structured argumentation by means of the ASPIC+ framework (Modgil and Prakken 2014). Arguments are used to support (or not) the passage of goals from one stage to the next and guide the passage of goals (diachronic support) and can be saved for future analysis (synchronic support). Besides, an argument can put together both the supporting beliefs and the supported goal in just one structure, facilitating future analysis. Arguments and their attack relations are part of argumentation frameworks (AF), which are generated in each stage to determine which goals pass to the next stage and which do not. To the best of our knowledge, this is the only formalization of this type of agent. Figure 1 shows the introduced goal reasoning model. It shows all the possible transitions of a goal from its active status, which can be seen as a desire in the BDI model until it becomes executive, which can seen as an intention in the BDI model. This depends on the arguments generated in each stage. Notice that in the life cycle of goals the cancelled status is also consid-

[1] AI is the acronym for Artificial Intelligence.

ered. Figure 2 shows the arguments generated for activating two goals and the attacks that arise between them. This figure was generated by the simulator ArgAgent (Jasinski, Morveli-Espinoza, and Tacla 2020), which was developed to evaluate our proposal.



Acceptable activation argument(s)
Acceptable evaluation argument(s)
DELIBERATION
Acceptable checking argument
Active
Pursuable
Chosen
Executive
Cancelled
2
1 or 3
2
1 or 3
4
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/or not longer preferred

Figure 1: Life cycle of goals.

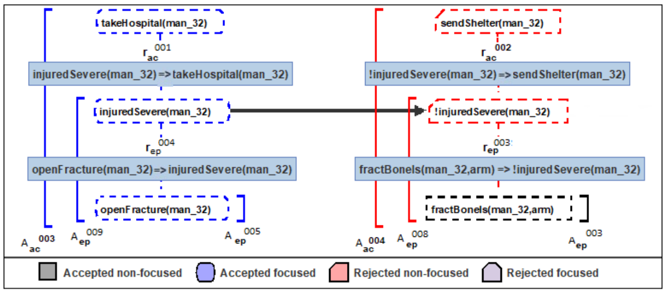
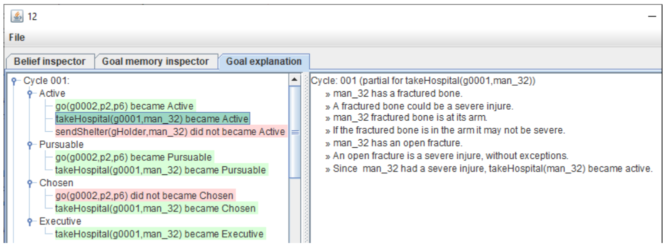


Figure 2: Arguments for goals $g_2 = take\_hospital(man\_32)$ and $g_3 = send\_shelter(man\_32)$.

The second contribution has to do with endowing BBGP-based agents with explainability abilities. Notice that both the synchronic and the diachronic support are related to explainability. For satisfying diachrony there should exist an argument that justifies the change in the status of goal and for satisfying synchrony, there should be a set of arguments, which represent the cognitive path. This path is saved in each AF, where arguments represent reasons in favor and against the change of the status of a goal. We generate complete and partial explanations depending on the arguments used to construct it; thus, an explanation is complete when the whole AF is used and is partial when an extension is used[2] of the AF. Figure 3 shows the partial explanation for query $\texttt{WHY}(g_2, ac)$[3] returned by ArgAgent as well. We can notice that it is generated in a very understandable language for humans. This is another benefit of using argumentation.

3 Results Discussion

This work presented an argumentation-based formalization for the BBGP model – which can be considered an extension of the BDI model – and an approach for explainable agency based on BBGP-based agents. The objective was that BBGP-based agents be able to explain their decision about the statuses of their goals. In order to achieve the objectives, we equipped BBGP-based agents with a structure and a mechanism to generate partial and complete explanations. We can say that due to the fine-granularity of BBGP model, it gives a good basement for constructing richer explanations. We demonstrate how our approach satisfies the desirable properties specified by (Castelfranchi and Paglieri 2007), that is, diachrony and synchrony. We first prove that the change of status of goals is always supported by arguments (diachrony) and that the agents save a cognitive path for explaining how a given goal reach out to its current status (synchrony).



```
Cycle 001:
  Active
    go(g0002,p2,p6) became Active
    takeHospital(g0001,man_32) became Active
    sendShelter(gHolder,man_32) did not became Active
  Pursuable
    go(g0002,p2,p6) became Pursuable
    takeHospital(g0001,man_32) became Pursuable
  Chosen
    go(g0002,p2,p6) did not became Chosen
    takeHospital(g0001,man_32) became Chosen
  Executive
    takeHospital(g0001,man_32) became Executive
```

```
Cycle: 001 (partial for takeHospital(g0001,man_32))
  » man_32 has a fractured bone.
  » A fractured bone could be a severe injure.
  » man_32 fractured bone is at its arm.
  » If the fractured bone is in the arm it may not be severe.
  » man_32 has an open fracture.
  » An open fracture is a severe injure, without exceptions.
  » Since  man_32 had a severe injure, takeHospital(man_32) became active.
```

Figure 3: Partial explanation for query $\texttt{WHY}(g_2, ac)$.

Acknowledgements

We thank CAPES/Brazil and CNPq Proc. 409523/2021-6.

References

Bratman, M. 1987. Intention, plans, and practical reason.

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

Guerrero, E.; Nieves, J. C.; and Lindgren, H. 2016. An activity-centric argumentation framework for assistive technology aimed at improving health. *Argument and Computation* 7(1):5–33.

Gunetti, P.; Thompson, H.; and Dodd, T. 2013. Autonomous mission management for uavs using soar intelligent agents. *International journal of systems science* 44(5):831–852.

Jasinski, H. M.; Morveli-Espinoza, M.; and Tacla, C. A. 2020. ArgAgent: a simulator of goal processing for argumentative agents. In *Computational Models of Argument (COMMA)*. IOS Press. 469–470.

Modgil, S., and Prakken, H. 2014. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation* 5(1):31–62.

Nieves, J. C., and Lindgren, H. 2014. Deliberative argumentation for service provision in smart environments. In *European Conference on Multi-Agent Systems*, 388–397. Springer.

Smuha, N. A. 2019. The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International* 20(4):97–106.

Winikoff, M.; Padgham, L.; Harland, J.; and Thangarajah, J. 2002. Declarative and procedural goals in intelligent agent systems. In *International Conference on Principles of Knowledge Representation and Reasoning*, 470–481.

[2] An extension in argumentation is a set of acceptable arguments, i.e. non-conflicting arguments.

[3] The meaning of the query is the following: Why goal $g_2 = take\_hospital(man\_32)$ was activated (denoted by ac)?