

# Modelling Human Mental-States in an Action Language following the Theory of Planned Behavior

Andreas Brännström, Juan Carlos Nieves

Umeå University, Department of Computing Science, SE-901 87, Umeå, Sweden

## Abstract

This paper introduces an action language for modelling the causality between a human's motivational beliefs and behavior in human activities. The action language is based on a psychological theory, the theory of planned behavior (TPB), which centers on three sets of beliefs that shape an individual's behavioral intentions: attitude, subjective norms, and perceived behavioral control. The language is modelled in the structure of a transition system in which states correspond to different mental states of an individual, and actions correspond to ways to transition between mental states in order to influence human intentions. We introduce its syntax and semantics. The semantics is characterized in terms of answer sets.

## Keywords

Human-aware planning, Action reasoning, Answer set programming, Theory of mind, Theory of planned behavior

## 1. Introduction

Human-aware planning is a way to improve the ability of autonomous systems to plan its actions in an environment populated and affected by humans [1]. A general problem in human-aware planning is to explain and model human behavior in order to predict future actions [2]. When an intelligent system's task, in addition, is to promote, encourage and change human behavior, it is of particular relevance that the underlying motivations and causes to behavior is explained.

Explaining human behavior is a difficult task due to its causal complexity [3], which can be analysed in multiple levels and abstractions. Computational models of human behavior have been based on, e.g., integrative physiology [4] on a low level, to, e.g., formalizations of social practices [5], i.e., typically performed behavior, on a high level. The present work aims to model human behavior from a personal, intermediate, level by looking at a human's motivation to behavior, and the human's underlying beliefs that shape motivation. The theory of planned behavior (TPB) [6] is a psychological theory for explaining an individual's motivation to engage in a behavior in a specific context. TPB focuses on three core motivational components: (1) an individual's attitude, e.g., expectations of the outcomes of a behavior, (2) subjective norms, e.g., an individual's perceived social pressure from others to behave in a certain manner, and (3)

---

*ASPOCP'21: 14th Workshop on Answer Set Programming and Other Computing Paradigms, September 20–27, 2021, Virtual*

✉ andreasb@cs.umu.se (A. Brännström); jcnieves@cs.umu.se (J. C. Nieves)

🌐 <https://www.umu.se/en/staff/andreas-brannstrom/> (A. Brännström);

<https://www.umu.se/en/staff/juan-carlos-nieves/> (J. C. Nieves)

🆔 0000-0001-9379-4281 (A. Brännström); 0000-0003-4072-8795 (J. C. Nieves)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

perceived behavioral control, e.g., an individual's self-efficacy in a behavior. TPB bases these three sources of motivation on the individual's particular beliefs in a moment, which together promote or demote engagement in the behavior.

A way for reasoning about human behavior is to look at it from the perspective of states and transitions between states. Action languages, such as  $\mathcal{A}$  [7] and  $\mathcal{C}_{TAID}$  [8], are declarative ways for specifying the behavior of transition systems, representing actions that result in transitions between states, i.e., configurations of environmental variables, in a domain. Such a domain can be a human mind, comprised of mental states including perceptions of the context/situation.

By following the concepts introduced in TPB, this work aims to formalize the causality between human beliefs and behavior in an action language. Such a formalization can be utilized by a rational agent to reason about the behavioral intentions of a human agent, and in addition reason about the rational agent's actions for changing a human agent's beliefs to influence motivations and intentions.

Modelling the causality between human beliefs and behavior involves identifying which beliefs and actions, out of many, that cause a behavior. This is a chain of direct and indirect effects that branch out to change mental states. This is a difficult task that involves the problem of ramification [9]. Furthermore, the priority of motivation must be taken into consideration, i.e., a human's current mental state determines which source of motivation that is most effective. The human agent's behavior must be determined through an indirect reasoning process where indicators to behavior are inferred through models of human reasoning. Given this action reasoning problem, and by considering components introduced in TPB, the following research questions arise:

- How can the causality between human motivation and behavior be expressed by an action reasoning language?
- How can an agent prioritise actions that cause change in the motivational beliefs in a human agent?

We approach these questions by introducing the action language  $\mathcal{C}_{TPB}$ . It is modelled in the structure of a transition system in which states correspond to different mental states of an individual and an environment, and actions correspond to ways to transition between states in order to influence human intentions. In order to characterize the semantics of the language we establish a link between  $\mathcal{C}_{TPB}$  and Answer Set Programming (ASP) [10]. This is done by presenting encodings of the language into logic programs following the answer set semantics. An ASP formalization, in addition to a formal characterization, gives correct implementations through solvers like clasp [11].

To capture human behavior, the action language  $\mathcal{C}_{TPB}$  has been developed in a knowledge elicitation process related to a case study [12] that deals with social behavior-change in children with autism. Together with experts in, e.g., psychology, we aim to develop principles of design to model human behavior.

The rest of this paper is organized as follows. First, we briefly present the theoretical framework; the theory of planned behavior. Syntax and semantics of the proposed action language is then presented which introduces an action reasoning approach to human-aware planning that utilizes the theory of planned behavior. The state-of-the-art in human-aware planning and agents modelling other agents is then presented and discussed. The paper is concluded by a discussion of the action language's potential, limitations, possible use-cases, and directions for future work.

## 2. Theoretical Framework and Background

This section explains the cognitive theory that has influenced the modelling of the proposed action reasoning framework for human-aware planning; Theory of Planned Behavior [6]. The results of a qualitative knowledge elicitation process is then presented that is captured in the semantics of the action language  $\mathcal{C}_{TPB}$ .

### 2.1. Theory of Planned Behavior

Theory of Planned Behavior (TPB) [6], is a cognitive theory for explaining and predicting an individual's intention to engage in a human activity at a specific time and place. The general idea is that the individual's beliefs about an activity shapes the individual's attitudes, subjective norm, and perceived behavioral control in the activity, which in turn promotes or inhibits engagement in the activity.

Attitude (A) refers to the degree to which an individual has a positive or negative evaluation of performing the activity. This entails a consideration of the outcomes of the activity. The overall attitude towards the activity is a consideration of each expected outcome of the activity quantified by the individual's valuation of that outcome.

Subjective norm (SN) refers to the individual's belief about whether people approve or disapprove of the activity. The overall subjective norms towards the activity is a consideration of each normative belief of the activity quantified by the individual's motivation to comply with that norm.

Perceived behavioral control (PBC) refers to an individual's perception of the ease or difficulty of performing the activity. The overall perceived behavioral control towards the activity is a consideration of each performance aspect of the activity quantified by the individual's perceived controllability of that performance aspect.

According to TPB, behavioral intention (BI) is the motivational factor that influences a given behavior, the likelihood that the human will initiate in an activity. Behavioral intention is the aggregation of the overall attitude, subjective norm, and perceived behavioral control. Specific activities can be more or less affected by these three predictors. Thus, an activity specific empirically derived weight is added to each predictor.

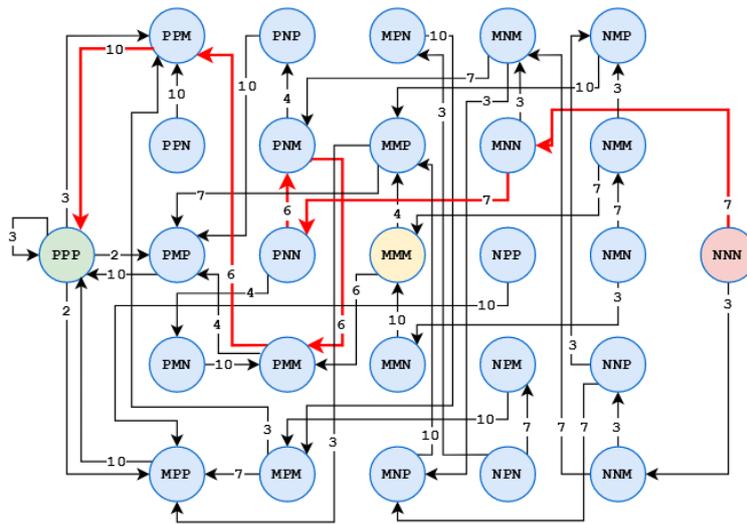
The following subsection presents a qualitative data collection process centered on the components of motivation introduced in TPB. Based on these components, mental states are specified which are captured in a particular transition system.

### 2.2. Knowledge elicitation results: Motivation decision-graph

This section presents the results of a knowledge elicitation process of a related use-case [12]. A qualitative study with 15 domain experts (psychologists, physiotherapists, occupational therapists, and special education teachers) was conducted. The study found that the most influential sources of motivation in a moment depends on an individual's current motivational beliefs in terms of TPB.

By assuming that the current mental state of an individual can be sufficiently recognized, i.e., the attitude (A), subjective norm (SN), and perceived behavioral control (PBC), and valued in

the scale *Negative* (N) (inhibiting behavior), *Medium* (M) (indifferent to behavior) and *Positive* (P) (promoting behavior), the participating experts were asked how a person's motivation to a behavior should be boosted depending on the person's current mental state. E.g., if the person currently has a mental state where A is negative, SN is negative and PBC is negative, then which aspect should be prioritised to boost? If so, what about a similar case where A is medium? By following this approach, 27 states are specified, each state consisting of the variables A, SN, or PBC, in which each variable has a value of negative, medium, or positive. Following the experts' suggestions, we can specify a state space with transition relations between states that represents a prioritised change of A, SN or PBC depending on the recognized mental state of the human. A mental state is denoted as A:SN:PBC, labeled by the value of each variable (e.g., PPP if A is Positive, SN is Positive and PBC is Positive). We can model these relations in terms of a weighted graph, called a motivation decision-graph (see Figure 1 [12] in which the red/bold path is the experts' prioritized path from state NNN to state PPP).



**Figure 1:** Motivation decision-graph.

By following the experts' suggestions of motivational focus in terms of the 27 mental states, a set of trends in the behavior of the system were found:

1. Prioritize to push attitude away from a negative state whenever possible, i.e., to a medium or positive state.
2. Prioritize to push subjective norm away from a negative state, if attitude is at least medium and control is positive.
3. Prioritize to push control away from a negative state if an internal or external motivator is already present, i.e., if attitude or subjective norm is at least medium.
4. If all three aspects are medium, prioritize to push attitude to a positive state.
5. The aim is to reach the state PPP, where all aspects are positive. If this state is reached,

then keep the variables steady or make small challenges in attitude, subjective norm, or control by pushing the state to any of the lower states MPP, PMP, or PPM.

These behaviors of the motivation decision-graph is formalized by the semantics of the action language  $\mathcal{C}_{TPB}$ . The language's syntax and semantics is presented in the following section.

### 3. Human Behavior in the Action Language $\mathcal{C}_{TPB}$

This section presents the syntax and semantics of the proposed action language,  $\mathcal{C}_{TPB}$ , following the results of a knowledge elicitation process presented in Section 2.2, in the light of the theory of planned behavior. Action descriptions in answer set semantics is then presented together with an implementation of the motivation decision-graph illustrated in Figure 1, in order to characterize the semantics of the  $\mathcal{C}_{TPB}$  action language.

#### 3.1. Human-Aware Transition Systems and Action Reasoning

The alphabet of  $\mathcal{C}_{TPB}$  consists of two nonempty disjoint sets of symbols  $\mathbf{F}$  and  $\mathbf{A}$ . They are called the set of fluents  $\mathbf{F}$  and the set of actions  $\mathbf{A}$ . A *fluent* expresses a property of an object in a world, and forms part of the description of states of this world. A *fluent literal* is a fluent or a fluent preceded by  $\neg$ . A *state*  $\sigma$  is a collection of fluents (informal approximation, see Definition 3). We say a fluent  $f$  holds in a state  $\sigma$  if  $f \in \sigma$ . We say a fluent literal  $\neg f$  holds in  $\sigma$  if  $f \notin \sigma$ .

**Definition 1 (Human-aware alphabet).** *Let  $\mathbf{A}$  be a non-empty set of actions and  $\mathbf{F}$  be a non-empty set of fluents.*

- $\mathbf{F} = \mathbf{F}^E \cup \mathbf{F}^H$  such that  $\mathbf{F}^E$  is a non-empty set of fluent literals describing observable items in an environment and  $\mathbf{F}^H$  is a non-empty set of fluent literals describing the mental-states of humans.  $\mathbf{F}^E$  and  $\mathbf{F}^H$  are pairwise disjoint.
- $\mathbf{F}^H = \mathbf{F}^A \cup \mathbf{F}^N \cup \mathbf{F}^C$  such that  $\mathbf{F}^A$ ,  $\mathbf{F}^N$  and  $\mathbf{F}^C$  are non-empty pairwise disjoint sets of fluent literals describing a human agent's attitude, subjective norm and perceived behavioral control, respectively.
- $\mathbf{A} = \mathbf{A}^E \cup \mathbf{A}^H$  such that  $\mathbf{A}^E$  is a non-empty set of actions that can be performed by a software agent and  $\mathbf{A}^H$  is non-empty set of actions that can be performed by a human agent.  $\mathbf{A}^E$  and  $\mathbf{A}^H$  are pairwise disjoint.

$\mathcal{C}_{TPB}$  is defined by three sub-languages: an action description language, an action observation language and an action query language.

**Definition 2.** *A human-aware domain description language  $D^h(\mathbf{A}, \mathbf{F})$  in  $\mathcal{C}_{TPB}$  consists of static and dynamic causal laws of the following form:*

- (*a* **causes**  $f_1, \dots, f_n$  **if**  $g_1, \dots, g_n$ ) (1)
- (*a* **influences attitude**  $f$  **if**  $f_1, \dots, f_n$ ) (2)
- (*a* **influences subjective norm**  $f$  **if**  $f_1, \dots, f_n$ ) (3)
- (*a* **influences control**  $f$  **if**  $f_1, \dots, f_n$ ) (4)
- ( $f_1, \dots, f_n$  **influences attitude**  $f$ ) (5)
- ( $f_1, \dots, f_n$  **influences subjective norm**  $f$ ) (6)
- ( $f_1, \dots, f_n$  **influences control**  $f$ ) (7)
- ( $f_1, \dots, f_n$  **if**  $g_1, \dots, g_m$ ) (8)
- ( $f_1, \dots, f_n$  **triggers**  $a$ ) (9)
- ( $f_1, \dots, f_n$  **allows**  $a$ ) (10)
- ( $f_1, \dots, f_n$  **inhibits**  $a$ ) (11)
- ( $f_1, \dots, f_n$  **promotes**  $a$ ) (12)
- ( $f_1, \dots, f_n$  **demotes**  $a$ ) (13)
- (**noconcurrency**  $a_1, \dots, a_n$ ) (14)
- (**default**  $g$ ) (15)

where  $a \in \mathbf{A}$  and  $a_i \in \mathbf{A}$  ( $0 \leq i \leq n$ ) and  $f_j \in \mathbf{F}$ , ( $0 \leq j \leq n$ ) and  $g_j \in \mathbf{F}$ , ( $0 \leq j \leq n$ ), and  $f \in \mathbf{F}^H$  then  $f$  is a ternary first order predicate, in which parameter one represents a mental state fluent  $MS \in \{Attitude, Subjective\_norm, Control\}$ , parameter two represents the MS's value  $V \in \{Negative, Medium, Positive\}$  and parameter three represents a point in time  $T$ , denoted as  $f(MS, V, T)$  or as  $f(\_, V, \_)$  when  $MS$  and  $T$  are explained by the context.

The semantics of a domain description  $D^h(A, F)$  is defined in terms of transition systems. An interpretation  $I$  over  $F$  is a complete and consistent set of fluents.

**Definition 3.** A state  $s \in S$  of the domain description  $D^h(A, F)$  is an interpretation over  $F$  such that

1. for every static causal law  $(f_1, \dots, f_n$  **if**  $g_1, \dots, g_n) \in D^h(A, F)$ , we have  $\{f_1, \dots, f_n\} \subseteq s$  whenever  $\{g_1, \dots, g_n\} \subseteq s$ .
2. for every static causal law  $(f_1, \dots, f_n$  **influences attitude**  $f) \in D^h(A, F)$ , we have  $\{f\} \subset s$  whenever  $\{f_1, \dots, f_n\} \subseteq s$ , and  
 $f \in F^A \wedge (f(\_, Positive, \_) \in s \vee f(\_, Medium, \_) \in s)$ , and  
 $(\exists f_i \in F^A (1 \leq i \leq n) \wedge f_i(\_, Negative, \_) \in s)$ , or  
 $\exists f_i \in F^A (1 \leq i \leq n) \wedge \exists f_j \in F^N (1 \leq j \leq n) \wedge \exists f_k \in F^C (1 \leq k \leq n) \wedge f_i(Medium) \in s \wedge f_j(Medium) \in s \wedge f_k(Medium) \in s$ .
3. for every static causal law  $(f_1, \dots, f_n$  **influences subjective norm**  $f) \in D^h(A, F)$ , we have  $\{f\} \subset s$  whenever  $\{f_1, \dots, f_n\} \subseteq s$ , and  
 $f \in F^N \wedge (f(\_, Positive, \_) \in s \vee f(\_, Medium, \_) \in s)$ , and  
 $\exists f_i \in F^A (1 \leq i \leq n) \wedge (f_i(\_, Medium, \_) \in s \vee f_i(\_, Positive, \_) \in s)$ , and  
 $\exists f_k \in F^C (1 \leq k \leq n) \wedge f_k(\_, Positive, \_) \in s$ .
4. for every static causal law  $(f_1, \dots, f_n$  **influences control**  $f) \in D^h(A, F)$ , we have  $\{f\} \subset s$  whenever  $\{f_1, \dots, f_n\} \subseteq s$ , and

$f \in F^C \wedge (f(\_, \text{Positive}, \_) \in s \vee f(\_, \text{Medium}, \_) \in s)$ , and  
 $\exists f_i \in F^A (1 \leq i \leq n) \wedge (f_i(\_, \text{Medium}, \_) \in s \vee f_i(\_, \text{Positive}, \_) \in s)$ , and  
 $\exists f_k \in F^C (1 \leq k \leq n) \wedge (f_k(\_, \text{Medium}, \_) \in s \vee f_k(\_, \text{Positive}, \_) \in s)$ .

$S$  denotes all the possible states of  $D^h(A, F)$ .

**Definition 4.** Let  $D^h(A, F)$  be a domain description and  $s$  a state of  $D^h(A, F)$ .

1. An inhibition rule ( $f_1, \dots, f_n$  inhibits  $a$ ) is active in  $s$ , if  $s \models f_1, \dots, f_n$ , otherwise the inhibition rule is passive. The set  $A_I(s)$  is the set of actions for which there exists at least one active inhibition rule in  $s$ .
2. A triggering rule ( $f_1, \dots, f_n$  triggers  $a$ ) is active in  $s$ , if  $s \models f_1, \dots, f_n$  and all inhibition rules of action  $a$  are passive in  $s$ , otherwise the triggering rule is passive in  $s$ . The set  $A_T(s)$  is the set of actions for which there exists at least one active triggering rule in  $s$ . The set  $\bar{A}_T(s)$  is the set of actions for which there exists at least one triggering rule and all triggering rules are passive in  $s$ .
3. An allowance rule ( $f_1, \dots, f_n$  allows  $a$ ) is active in  $s$ , if  $s \models f_1, \dots, f_n$  and all inhibition rules of action  $a$  are passive in  $s$ , otherwise the allowance rule is passive in  $s$ . The set  $A_A(s)$  is the set of actions for which there exists at least one active allowance rule in  $s$ . The set  $\bar{A}_A(s)$  is the set of actions for which there exists at least one allowance rule and all allowance rules are passive in  $s$ .
4. A promoting rule ( $f_1, \dots, f_n$  promotes  $a$ ) is active in  $s$ , if  $a \in \mathbf{A}^H$  and  $s \models f_1, \dots, f_n$  and all inhibition rules and demoting rules of action  $a$  are passive in  $s$ , otherwise the promoting rule is passive in  $s$ . The set  $A_P(s)$  is the set of actions for which there exists at least one active promoting rule in  $s$ . The set  $\bar{A}_P(s)$  is the set of actions for which there exists at least one promoting rule and all promoting rules are passive in  $s$ .
5. A demoting rule ( $f_1, \dots, f_n$  demotes  $a$ ) is active in  $s$ , if  $a \in \mathbf{A}^H$  and  $s \models f_1, \dots, f_n$  and all inhibition rules and promoting rules of action  $a$  are passive in  $s$ , otherwise the demoting rule is passive in  $s$ . The set  $A_D(s)$  is the set of actions for which there exists at least one active demoting rule in  $s$ . The set  $\bar{A}_D(s)$  is the set of actions for which there exists at least one demoting rule and all demoting rules are passive in  $s$ .
6. A dynamic causal law ( $a$  causes  $f_1, \dots, f_n$  if  $g_1, \dots, g_n$ ) is applicable in  $s$ , if  $s \models g_1, \dots, g_n$ .
7. A static causal law ( $f_1, \dots, f_n$  if  $g_1, \dots, g_n$ ) is applicable in  $s$ , if  $s \models g_1, \dots, g_n$ .
8. A dynamic causal law ( $a$  influences attitude  $f$  if  $f_1, \dots, f_n$ ) is applicable in  $s$ , if  $s \models f_1, \dots, f_n$ , and  $f \in F^A$ , and  $\exists f_i \in F^A (1 \leq i \leq n)$ , and  $\exists f_j \in F^N (1 \leq j \leq n)$ , and  $\exists f_k \in F^C (1 \leq k \leq n)$ .
9. A dynamic causal law ( $a$  influences subjective norm  $f$  if  $f_1, \dots, f_n$ ) is applicable in  $s$ , if  $s \models f_1, \dots, f_n$ , and  $f \in F^N$ , and  $\exists f_i \in F^A (1 \leq i \leq n)$ , and  $\exists f_j \in F^N (1 \leq j \leq n)$ , and  $\exists f_k \in F^C (1 \leq k \leq n)$ .
10. A dynamic causal law ( $a$  influences control  $f$  if  $f_1, \dots, f_n$ ) is applicable in  $s$ , if  $s \models f_1, \dots, f_n$ , and  $f \in F^C$ , and  $\exists f_i \in F^A (1 \leq i \leq n)$ , and  $\exists f_j \in F^N (1 \leq j \leq n)$ , and  $\exists f_k \in F^C (1 \leq k \leq n)$ .

11. A static causal law  $(f_1, \dots, f_n \text{ influences attitude } f)$  is applicable in  $s$ , if  $s \models f_1, \dots, f_n$ , and  $f \in F^A$ , and  $\exists f_i \in F^A (1 \leq i \leq n)$ , and  $\exists f_j \in F^N (1 \leq j \leq n)$ , and  $\exists f_k \in F^C (1 \leq k \leq n)$ .
12. A static causal law  $(f_1, \dots, f_n \text{ influences subjective norm } f)$  is applicable in  $s$ , if  $s \models f_1, \dots, f_n$ , and  $f \in F^N$ , and  $\exists f_i \in F^A (1 \leq i \leq n)$ , and  $\exists f_j \in F^N (1 \leq j \leq n)$ , and  $\exists f_k \in F^C (1 \leq k \leq n)$ .
13. A static causal law  $(f_1, \dots, f_n \text{ influences control } f)$  is applicable in  $s$ , if  $s \models f_1, \dots, f_n$ , and  $f \in F^C$ , and  $\exists f_i \in F^A (1 \leq i \leq n)$ , and  $\exists f_j \in F^N (1 \leq j \leq n)$ , and  $\exists f_k \in F^C (1 \leq k \leq n)$ .

**Definition 5 (Trajectory).** Let  $D^h(A, F)$  be a domain description. A trajectory  $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$  of  $D^h(A, F)$  is a sequence of sets of actions  $A_i \subseteq A$  and states  $s_i$  of  $D^h(A, F)$  satisfying the following conditions for  $0 \leq i < n$ :

1.  $(s_i, A, s_{i+1}) \in S \times 2^A \setminus \{\emptyset\} \times S$
2.  $A_T(s_i) \subseteq A_{i+1}$
3.  $A_P(s_i) \subseteq A_{i+1}$
4.  $A_D(s_i) \subseteq A_{i+1}$
5.  $\overline{A}_T(s_i) \cap A_{i+1} = \emptyset$
6.  $\overline{A}_A(s_i) \cap A_{i+1} = \emptyset$
7.  $\overline{A}_I(s_i) \cap A_{i+1} = \emptyset$
8.  $\overline{A}_P(s_i) \cap A_{i+1} = \emptyset$
9.  $\overline{A}_D(s_i) \cap A_{i+1} = \emptyset$
10.  $|A_i \cap B| \leq 1$  for all (noconcurrency  $B \in D^h(A, F)$ ).

**Definition 6.** The action observation language of  $\mathcal{C}_{TPB}$  consists of expressions of the following form:

$$(f \text{ at } t_i) \quad (a \text{ occurs\_at } t_i) \quad (16)$$

where  $f \in \mathbf{F}$ ,  $a$  is an action and  $t_i$  is a point of time.

**Definition 7 (Action Theory).** Let  $D^h$  be a domain description and  $O$  be a set of observations. The pair  $(D^h, O)$  is called an action theory.

**Definition 8 (Trajectory Model).** Let  $(D^h, O)$  be an action theory. A trajectory  $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$  of  $D^h$  is a trajectory model of  $(D^h, O)$ , if it satisfies all observations of  $O$  in the following way:

1. if  $(f \text{ at } t) \in O$ , then  $f \in s_t$
2. if  $(a \text{ occurs\_at } t) \in O$ , then  $a \in A_{t+1}$ .

Let us observe that given a trajectory  $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$  where  $A_i \subseteq \mathbf{A}$  ( $0 \leq i \leq n$ ) and  $\mathbf{A}$  is a set of actions that can be performed either by a human-agent or a software planner-agent. Actions performed by a human-agent can be movement between areas in the environment, while actions by a software planner-agent are adaptations of the environment that indirectly influences a human-agent's mental-state fluents, i.e., attitude, subjective norm and control.

**Definition 9 (Action Query Language).** *The action query language of  $\mathcal{C}_{TPB}$  regards assertions about executing sequences of actions with expressions that constitute trajectories. A query is of the following form:*

$(f_1, \dots, f_n \text{ after } A_i \text{ occurs\_at } t_i, \dots, A_m \text{ occurs\_at } t_m)$   
*where  $f_1, \dots, f_n$  are fluent literals  $\in \mathbf{F}^E \cup \mathbf{F}^H$ ,  $A_i, \dots, A_m$  are sub-sets of  $\mathbf{A}^E \cup \mathbf{A}^H$ , and  $t_i, \dots, t_m$  are points in time.*

### 3.2. Action Descriptions in Answer Set Semantics

This section presents translations in answer set programs of the expressions as part of  $\mathcal{C}_{TPB}$ . These expressions incorporate the result from the knowledge elicitation process, specified to capture explanations of human behavior and behavior-change. Descriptions of expression 1, 9, 10, 11, 14, and 15 follows the definitions in [8].

In the following translations, the value of a mental state variable, i.e., attitude, subjective norm and control, is represented by  $h \in \{Positive, Medium, Negative\}$  in a time point  $T$ . Actions are coupled with a planner agent (*agent*) or a human agent (*human*).

$(a \text{ influences attitude } h \text{ if } f_1, \dots, f_n)$  (2)

$attitude(h, T+1) : - holds(f_1, T), \dots, holds(f_n, T), fluent(f_1), \dots, fluent(f_n),$   
 $occurs(a, T), action(a, agent), time(T), T < n, \#int(T), motivate(attitude, h, T).$

$(a \text{ influences subjective norm } h \text{ if } f_1, \dots, f_m)$  (3)

$subjective\_norm(h, T+1) : - holds(f_1, T), \dots, holds(f_n, T), fluent(f_1), \dots, fluent(f_n),$   
 $occurs(a, T), action(a, agent), time(T), T < n, \#int(T), motivate(subjective\_norm, h, T).$

$(a \text{ influences control } h \text{ if } f_1, \dots, f_m)$  (4)

$control(h, T+1) : - holds(f_1, T), \dots, holds(f_n, T), fluent(f_1), \dots, fluent(f_n),$   
 $occurs(a, T), action(a, agent), time(T), T < n, \#int(T), motivate(control, h, T).$

$(f_1, \dots, f_n \text{ influences attitude } h)$  (5)

$attitude(h, T+1) : - holds(f_1, T), \dots, holds(f_n, T), fluent(f_1), \dots, fluent(f_n),$   
 $time(T), T < n, \#int(T), motivate(attitude, h, T).$

$(f_1, \dots, f_n \text{ influences subjective norm } h)$  (6)

$subjective\_norm(h, T+1) : - holds(f_1, T), \dots, holds(f_n, T), fluent(f_1), \dots, fluent(f_n),$   
 $time(T), T < n, \#int(T), motivate(subjective\_norm, h, T).$

$(f_1, \dots, f_n \text{ influences control } h)$  (7)

$control(h, T+1) : - holds(f_1, T), \dots, holds(f_n, T), fluent(f_1), \dots, fluent(f_n),$   
 $time(T), T < n, \#int(T), motivate(control, h, T).$

$(f_1, \dots, f_n \text{ promotes } a)$  (12)

$holds(occurs(a), T+1) : - not holds(ab(occurs(a)), T+1), holds(f_1, T), \dots, holds(f_n, T),$   
 $fluent(f_1), \dots, fluent(f_n), action(a, human), time(T), T < n, \#int(T).$

$(f_1, \dots, f_n \text{ demotes } a)$  (13)

$holds(ab(occurs(a)), T+1) : - holds(f_1, T), \dots, holds(f_n, T), fluent(f_1), \dots, fluent(f_n),$   
 $action(a, human), time(T), T < n, \#int(T).$

The above expressions can be modelled according to a specific domain, to capture beliefs of an environment and their influence on human behavior. Accompanying each expression, a domain independent set of rules are applied to filter out solution candidates in a generated trajectory. This is expressed by the predicate *motivate* which adds a set of integrity constraints based on the motivation decision-graph (presented in Subsection 2.2), defined below.

**Definition 10.** A motivation decision-graph MDG is a transition system that is a tuple of the form  $MDG = (M, Act, T, O, MP, L)$  where  $M$  is a non-empty set of states, denoting mental states of a human agent,  $Act$  is a set of actions,  $T \subseteq M \times M$  is a non-empty set of transition relations denoting legal transitions between mental states,  $O$  is a set of initial states,  $MP$  a set of atomic propositions, and  $L$  is a function that defines which propositions  $\in MP$  valid in each state  $\in M$ .

In the following definition, we introduce a logic program that characterize the motivation decision-graph of Figure 1. This logic program will help to define the semantics of  $\mathcal{C}_{TPB}$ .

**Definition 11.** Let  $P_{MDG}$  be the following logic program:

```

1  value(1..3). % 1: Negative, 2: Medium, 3: Positive.
2  mind(attitude). mind(norm). mind(control).
3
4  init_on(attitude,1). init_on(norm,1). init_on(control,1).
5  goal_on(attitude,3). goal_on(norm,3). goal_on(control,3).
6  plan_length(6). #show motivate/3.
7  { motivate(MS,V,T) : mind(MS), value(V) } = 1 :- plan_length(M), T = 1..M.
8
9  motivate(MS,T) :- motivate(MS,_,T).
10 on(MS,V,0) :- init_on(MS,V).
11 on(MS,V,T) :- motivate(MS,V,T).
12 on(MS,V,T+1) :- on(MS,V,T), not motivate(MS,T+1), not plan_length(T).
13
14 blocked(MS,V-1,T+1) :- on(MS,V,T), not plan_length(T).
15 blocked(MS,V-1,T) :- blocked(MS,V,T), value(V).
16
17 % C-TPB: Integrity constraints
18 :- motivate(MS,V,T), blocked(MS,V,T).
19 :- motivate(MS,T), on(MS,V,T-1), blocked(MS,V,T).
20 :- goal_on(MS,V), not on(MS,V,M), plan_length(M).
21 :- { on(MS,V,T) } != 1, mind(MS), plan_length(M), T = 1..M.
22
23 % Restrict Attitude
24 :- motivate(attitude, 1, T+1), on(attitude, 1, T).
25 :- motivate(attitude, 2, T+1), on(attitude, 2, T).
26 :- motivate(attitude, 3, T+1), on(attitude, 3, T).
27 :- motivate(attitude, 3, T+1), on(attitude, 1, T).
28 :- motivate(attitude, 1, T+1), on(attitude, 3, T).
29 :- motivate(attitude, 2, T+1), on(attitude, 3, T).
30
31 % Restrict Norm
32 :- motivate(norm, 1, T+1), on(norm, 1, T).
33 :- motivate(norm, 2, T+1), on(norm, 2, T).
34 :- motivate(norm, 3, T+1), on(norm, 3, T).

```

```

35 :- motivate(norm, 3, T+1), on(norm, 1, T).
36 :- motivate(norm, 1, T+1), on(norm, 3, T).
37 :- motivate(norm, 2, T+1), on(norm, 3, T).
38
39 % Restrict Control
40 :- motivate(control, 1, T+1), on(control, 1, T).
41 :- motivate(control, 2, T+1), on(control, 2, T).
42 :- motivate(control, 3, T+1), on(control, 3, T).
43 :- motivate(control, 3, T+1), on(control, 1, T).
44 :- motivate(control, 1, T+1), on(control, 3, T).
45 :- motivate(control, 2, T+1), on(control, 3, T).
46
47 % Push attitude whenever possible to a medium or positive state.
48 :- motivate(norm, 2, T+1), on(attitude, 1, T).
49 :- motivate(norm, 3, T+1), on(attitude, 1, T).
50 :- motivate(norm, 2, T+1), on(attitude, 2, T).
51 :- motivate(control, 2, T+1), on(attitude, 1, T).
52 :- motivate(control, 3, T+1), on(attitude, 1, T).
53 :- motivate(control, 2, T+1), on(attitude, 2, T).
54
55 % Push subjective norm if attitude is at least medium and control is positive.
56 % This constraint is mostly covered above, extended with:
57 :- motivate(control, 3, T+1), on(norm, 2, T).
58
59 % Push control if attitude or subjective norm is at least medium.
60 % This constraint is mostly covered above, extended with:
61 :- motivate(norm, 3, T+1), on(control, 1, T).

```

The above logic program ran in Clingo 4.5.4 generates trajectories following the motivational transition-graph in Figure 1. For instance, one get the following trajectory:

```

motivate(attitude,2,1), motivate(attitude,3,2), motivate(control,2,3),
motivate(norm,2,4), motivate(norm,3,5), motivate(control,3,6)

```

The first parameter in motivate/3 corresponds to a mental state fluent, e.g., Attitude. The second parameter in motivate/3 corresponds to the value of a mental state fluent, where value(1) equals to *Negative*, value(2) equals to *Medium*, and value(3) equals to *Positive*. The third parameter in motivate/3 corresponds to a point in time. The motivation decision-graph, prototyped in the above logic program (shared and openly accessible online<sup>1</sup>), works as domain-independent heuristics for any domain-specific logic program following the  $\mathcal{C}_{TPB}$  action language.

In order to define the semantics of  $\mathcal{C}_{TPB}$ , we characterize trajectory models in terms of answer sets. This is formalized by the following theorem:

**Theorem 1.** *Let  $(D^h, O_{initial})$  be an action theory such that  $O_{initial}$  are the fluent observations of the dynamic environment in the current state, and the fluents of the currently estimated mental state of the human (Attitude, Subjective norm, and Control). Let  $Q$  be a query, according to Definition 9 and let*

$$A_Q = \{(a \text{ occurs\_at } t_i) \mid a \in A_i, 1 \leq i \leq m\}.$$

<sup>1</sup>Source of the motivation decision-graph prototype: <https://git.io/Ju9vh>

Let  $\mathcal{T}$  denote the translation of  $\mathcal{C}_{TPB}$  into a logic program according to the mapping introduced by Section 3.2.

Then, the following statements hold true:

1. If there is a trajectory model  $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_m \rangle$  where  $A_i \subseteq \mathbf{A}$  ( $0 \leq i \leq m$ ) of  $\mathcal{C}_{TPB}(D^h, O_{initial} \cup A_Q)$ , then there is an answer set  $\mathcal{A}$  of logic program  $\mathcal{T}(\mathcal{C}_{TPB}(D^h, O_{initial} \cup A_Q) \cup P_{MDG})$ , such that for all fluents  $f \in \mathbf{F}^E \cup \mathbf{F}^H$  at the time points  $0 \leq k \leq m$

- (a)  $holds(f, k) \in \mathcal{A}$ , if  $s_k \models f$ ,
- (b)  $holds(neg(f), k) \in \mathcal{A}$ , if  $s_k \models \neg f$ .
- (c)  $holds(occurs(a), k) \in \mathcal{A}$ , if  $a \in A_{k+1}$
- (d)  $holds(neg(occurs(a)), k) \in \mathcal{A}$ , if  $a \notin A_{k+1}$

2. If there is an answer set  $\mathcal{A}$  of a program  $\mathcal{T}(\mathcal{C}_{TPB}(D^h, O_{initial} \cup A_Q) \cup P_{MDG})$  and at time point  $0 \leq k \leq m$

- (a)  $s_k = \{f \mid holds(f, k) \in \mathcal{A}\} \cup \{\neg f \mid holds(neg(f), k) \in \mathcal{A}\}$
- (b)  $A_{k+1} = \{a \mid holds(occurs(a), k) \in \mathcal{A}\}$

then there is a trajectory model  $\langle s_0, A_1, s_1, A_2, \dots, A_m, s_m \rangle$  of  $\mathcal{C}_{TPB}(D^h, O_{initial} \cup A_Q)$ .

**Proof 1.** (Sketch) Let us start by observing that  $\mathcal{C}_{TAID}$  [8] is a sub-language of  $\mathcal{C}_{TPB}$ . Hence, giving a  $\mathcal{C}_{TPB}$  program  $P_{TPB}$  there is a  $\mathcal{C}_{TAID}$  program  $P_{TAID}$  such that  $P_{TAID} \subseteq P_{TPB}$ .

Let  $\mathcal{T}_{TAID}$  be the translation of  $P_{TAID}$  into a logic program according to [8], and, let  $\mathcal{T}_{TPB}$  denote the translation of  $P_{TPB}$  into a logic program according to the mapping introduced by Section 3.2. One can see that  $\mathcal{T}_{TAID}$  is a subset of  $\mathcal{T}_{TPB}$ . One can observe that: Since  $P_{MDG}$  is only inferring paths in a graph, then if  $\mathcal{A} \cup A$  is an answer set of  $\mathcal{T}_{TPB}$  such that  $\mathcal{A} \subseteq \mathcal{L}_{\mathcal{T}_{TAID}}$  then  $\mathcal{A}$  is an answer set of  $P_{TAID}$ . Since Axioms 2-7 are similar to axiom 8 in Definition 2 and the axioms 12-13 are similar to axiom 9 in Definition 2, the proof follows by Theorem 1 in [8].

### 3.3. Case study: Promote social behavior in children with autism

This subsection exemplifies a use-case of  $\mathcal{C}_{TPB}$ . Children with autism commonly experience difficulties in social situations. Lights, sounds, people, and lack of guidance can result in stress and avoiding behavior. Let us consider an interactive virtual reality (VR) aid for children with autism for practicing social situations [13]. A rational agent is embedded in the application, which has the ability to adapt the virtual environment in order to promote the child to explore the scenario. The agent models the human using  $\mathcal{C}_{TPB}$ . A domain description  $D^h(A, F)$  in  $\mathcal{C}_{TPB}$  can include:  $\mathbf{A}^E = \{\text{increase\_sound}(L), \text{decrease\_sound}(L), \text{increase\_light}(L), \text{decrease\_light}(L), \text{increase\_guides}(L), \text{decrease\_guides}(L), \text{increase\_people}(L), \text{decrease\_people}(L)\}$ .  $\mathbf{A}^H = \{\text{move}(\text{Position})\}$ .  $\mathbf{F}^E = \{\text{sound}(L), \text{light}(L), \text{people}(L), \text{guides}(L)\}$ .  $\mathbf{F}^A = \{\text{attitude}(V)\}$ .  $\mathbf{F}^N = \{\text{norm}(V)\}$ .  $\mathbf{F}^C = \{\text{control}(V)\}$ .  $L \in \{\text{Low}, \text{Medium}, \text{High}\}$ .  $V \in \{\text{Negative}, \text{Medium}, \text{Positive}\}$ . A set of causal laws is declared, such as:  $\text{increase\_guides}(\text{High})$  influences  $\text{control}(\text{Medium})$  if  $\text{control}(\text{Low})$ . Based on an initial environment state (e.g.,  $\text{people}(\text{High}), \text{light}(\text{High}), \text{sound}(\text{High}), \text{guides}(\text{Low})$ ) and an initial mental state (e.g.,  $\text{attitude}(\text{Medium}), \text{norm}(\text{Negative}), \text{control}(\text{Negative})$ ), adaptation-plans are generated for promoting human actions. For instance, the following plan:  $\text{increase\_guides}(\text{High}); \text{decrease\_light}(\text{Medium}); \text{decrease\_sound}(\text{Medium})$ .

## 4. Discussion and Related Work

Previously, we have done work on context-reasoning (through Activity Theory [14, 15]) and now our work deals with the interaction process (through human-aware planning). The introduced issues of modelling mental states of a human in an action language have not been explored so far by the community of formal action reasoning. The proposed action reasoning language  $\mathcal{C}_{TPB}$  provides a human-aware alphabet for describing motivational aspects in an environment and the mental-states of humans, as well as actions by which variables of the mental state or the environment directly or indirectly can be changed.

There is a diverse body of research related to the ideas presented in the current work [16, 17, 18]. For instance, plan recognition as planning, originally introduced by Ramirez and Geffner [19], use planning algorithms to enable an agent to recognize the goals and plans of other agents. A related line of research introduces Empathetic planning [16]. In their work, empathy is defined as the ability to understand and share the thoughts and feelings of another. Following this definition, an assistive empathetic agent is formalized able to reason about the preferences of an empathizee [16]. The current work can advance the state-of-the-art of empathic agents by using a formalization of a psychological theory, the theory of planned behavior (TPB), in an attempt to model human beliefs and motivation.

Usually, when we talk about theory of mind, i.e., agents modelling other agents [20], we say that the agent has beliefs about the human's beliefs, but what does the human believe? In the  $\mathcal{C}_{TPB}$  action language, the agent can reason about particular beliefs about the human's beliefs, i.e., the human's attitude, subjective norm and perceived behavioral control in a specific behavior. In this way, the agent's theory of the mind of the human is particular and concise, making it possible to deliberate about causes to an individual's behavior, and how to promote human actions.

## 5. Conclusion and Future Work

We have introduced the action language  $\mathcal{C}_{TPB}$  based on a psychological theory, the theory of planned behavior, explaining a human's intention to engage in a human activity, and presented how the language can be applied to represent and reason about actions for altering mental-states to promote behavior. By utilizing the action language, an agent can reason about specific human beliefs, i.e., the human's attitude, subjective norm and perceived behavioral control in a situation. In this way, the agent acquires a particular theory of the mind of the human to deliberate about a human agent's intentions. On a low level, the language captures a human agent's beliefs about the environment and how these beliefs correspond to mental states. On a high level, the language captures a suitable direction of motivation based on a human's current mental state. In this way, a priority of motivation can be utilized for picking the most suitable actions to alterate the environment in order to change the human's motivational beliefs and influence human behavior. Future work concerns incorporating expressions dealing with probability distributions in human behavior into the language. Furthermore, we aim to explore ways to incorporate customized decision-graphs into the semantics of  $\mathcal{C}_{TPB}$ , thus going beyond hardwired transition rules. For instance, we aim to develop an emotion decision-graph which extends the semantics of  $\mathcal{C}_{TPB}$  with models to reason about human emotions and emotional-change.

## References

- [1] T. Chakraborti, Foundations of Human-Aware Planning-A Tale of Three Models, Ph.D. thesis, Arizona State University, 2018.
- [2] M. Cirillo, L. Karlsson, A. Saffiotti, Human-aware task planning: An application to mobile robots, *ACM Transactions on Intelligent Systems and Technology (TIST)* 1 (2010) 1–26.
- [3] F. I. Dretske, *Explaining behavior: Reasons in a world of causes*, MIT press, 1991.
- [4] D. Fass, R. Lieber, Rationale for human modelling in human in the loop systems design, in: 2009 3rd Annual IEEE Systems Conference, IEEE, 2009, pp. 27–30.
- [5] F. Dignum, Interactions as social practices: towards a formalization, *arXiv preprint arXiv:1809.08751* (2018).
- [6] I. Ajzen, et al., The theory of planned behavior, *Organizational behavior and human decision processes* 50 (1991) 179–211.
- [7] M. Gelfond, V. Lifschitz, *Action languages* (1998).
- [8] S. Dworschak, S. Grell, V. Nikiforova, T. Schaub, J. Selbig, Modeling Biological Networks by Action Languages via Answer Set Programming, *Constraints* 13 (2008) 21–65.
- [9] L. Giordano, A. Martelli, C. Schwind, Ramification and causality in a modal action logic, *Journal of logic and computation* 10 (2000) 625–662.
- [10] V. Lifschitz, *Answer set programming*, Springer International Publishing, 2019.
- [11] M. Gebser, R. Kaminski, B. Kaufmann, M. Ostrowski, T. Schaub, S. Thiele, *A user’s guide to gringo, clasp, clingo, and iclingo* (2008).
- [12] A. Brännström, J. C. Nieves, T. Kampik, E. Domellöf, L. Gu, M. Liljeström, Human-aware planning in virtual reality for facilitating social behavior in autism (2021). Manuscript submitted for publication.
- [13] B. Andreas, T. Kampik, J. C. Nieves, Towards human-aware epistemic planning for promoting behavior-change, in: *Workshop on Epistemic Planning (EpiP)@ ICAPS*, Online, October 26-30, 2020, 2020.
- [14] E. Guerrero, J. C. Nieves, H. Lindgren, An activity-centric argumentation framework for assistive technology aimed at improving health, *Argument & Computation* 7 (2016) 5–33.
- [15] J. Oetsch, J.-C. Nieves, A knowledge representation perspective on activity theory, *arXiv preprint arXiv:1811.05815* (2018).
- [16] M. Shvo, S. A. McIlraith, Towards empathetic planning, *arXiv preprint arXiv:1906.06436* (2019).
- [17] J. Blount, M. Gelfond, Reasoning about the intentions of agents, in: *Logic Programs, Norms and Action*, Springer, 2012, pp. 147–171.
- [18] A. Gabaldon, Activity recognition with intended actions, in: *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.
- [19] M. Ramirez, H. Geffner, Plan recognition as planning, in: *Proceedings of the 21st international joint conference on Artificial intelligence*. Morgan Kaufmann Publishers Inc, 2009, pp. 1778–1783.
- [20] S. V. Albrecht, P. Stone, Autonomous agents modelling other agents: A comprehensive survey and open problems, *Artificial Intelligence* 258 (2018) 66–95.