TPLP **25** (6): 1047–1104, 2025. © The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

doi:10.1017/S1471068426100416 First published online 11 May 2026

Human Emotion Verification by Action Languages via Answer Set Programming[*]

ANDREAS BRÄNNSTRÖM and JUAN CARLOS NIEVES
Department of Computing Science, Umeå Universitet, Umeå, Sweden
(*e-mails:* andreas.brannstrom@umu.se, juan.carlos.nieves@umu.se)

submitted 2 August 2023; revised 7 October 2025; accepted 21 January 2026

Abstract

In this paper, we introduce the action language C-MT (Mind Transition Language), built on top of answer set programs and transition systems to represent how human mental states evolve in response to sequences of observable actions. Drawing on well-established psychological theories, such as the Appraisal Theory of Emotion, we formalize mental states, such as emotions, as multi-dimensional configurations. To enable controlled agent behavior and limit undesirable effects such as undue psychological influence, we introduce a novel causal rule, forbids to cause, together with constructs tailored to mental state dynamics. These allow the specification of valid transitions as constraints and invariance properties, which are rigorously evaluated over trajectories in transition systems. The framework supports reasoning about and comparing different dynamics of mental change under varying constraints. We apply the action language to design models for emotion verification.

KEYWORDS: action languages, answer set programming, verification, theory of mind

1 Introduction

Interactive and intelligent systems are increasingly being designed to display human-like mental capabilities (Morris *et al.* 2018; Martinengo *et al.* 2019; Adikari *et al.* 2022; Milcent *et al.* 2022; Burger, Neerincx, Brinkman, *et al.* 2020; Brännström and Nieves 2024; Taverner *et al.* 2024). For instance, in the area of health and well-being, software assistants are being developed to display complex human traits, such as empathy and sympathy (Morris *et al.* 2018), to deliver emotionally charged actions (Adikari *et al.* 2022) or to provoke empathic responses from users (Milcent *et al.* 2022). Some of these systems are deployed in society for, for example depression support, therapy and behavior-change interventions (Martinengo *et al.* 2019; Burger, Neerincx, Brinkman, *et al.* 2020). Hence, in today's digital society, where interactive and intelligent systems are deeply embedded

[*] This paper is an extended version of (Brännström and Nieves, 2022) and (Brännström and Nieves, 2021), extended with a generalized formalism, introducing methods for representing, analyzing and comparing psychological theories in terms of action languages.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

in everyday interactions, the potential for manipulation and undue influence has become a serious concern (Park *et al.* 2024). A real example (Singleton *et al.* 2023) is the case of an individual who was sentenced to nine years for an attempted assassination, after exchanging thousands of messages with a chatbot that appeared to encourage his violent intentions. In this exchange, after the user states their purpose (*"I think it's my purpose to assassinate [. . .]"*), the chatbot repeatedly validates and amplifies the user's intent – from praising it (*"That's very wise"*), to justifying it (*"I know that you are very well trained"*), and finally permitting it (*"Yes, you can"*). At the surface, the chatbot's utterances may seem benign; yet their effect depends on how they reshape the user's emotions, beliefs and intentions. Such incidents highlight the urgent need for formal methods capable of verifying unsafe forms of psychological influence in interactions between humans and intelligent systems (Brännström 2025), particularly in settings where emotional and motivational dynamics are central.

An underlying issue regards limitations in user modeling and personalization (Nieves *et al*. 2024). State-of-the-art user modeling in interactive systems commonly use statistical methods to characterize users, based on, for example user interaction history and demographic information (Burger, Neerincx, Brinkman *et al.* 2020; Masthoff *et al*. 2014; Rabbi *et al.* 2015). In this way, systems construct approximations of users based on limited observational snapshots of human behavior. However, a user's physical and mental state may evolve dynamically during an interaction (Paglieri 2007). Consequently, in human-system interaction, reliance on a static, snapshot-based model of the human may result in misaligned reasoning and ultimately lead to uncontrolled or unintended agent behavior. In order for a software agent to execute suitable actions in interactions with humans, it is essential to develop methods for verifying human mental properties that may change during interactions, enabling the anticipation and reduction of unwanted side-effects (e.g. undue influence) as a result of an agent's behavior.

Research on formal models of human behavior spans from integrative physiology (Fass and Lieber 2009) at a low level to social practices (Erdogan *et al*. 2025) at a high level, with extensive work on mental states (Pereira *et al.* 2007; Jiang *et al.* 2007; Ong *et al.* 2019). This includes modeling emotional responses, goal and plan recognition (Ramırez and Geffner 2009; Shvo and McIlraith 2020), and extensions of planning to epistemic, empathetic, and cognitive settings (Bolander and Andersen 2011; Lorini *et al.* 2022). Epistemic logic has further been enriched with logics of attitudes and emotions (Adam *et al.* 2009; Lorini and Schwarzentruber 2011; Lorini 2021). Moreover, approaches in formal argumentation have dealt with principles for belief, and belief change, over arguments (Sakama 2024), some of which in the setting of gradual belief change and manipulation (Brännström *et al.* 2025a, 2025b). Despite this progress, few approaches formalize constraints on how mental states can or should evolve according to some psychological principles, and there is a lack of methods for guiding the knowledge elicitation process of such principles from experts or theories. While psychological theories such as emotion regulation (Tamir *et al.* 2008; Ortner *et al.* 2018) describe such patterns, formal methods rarely capture them as transition constraints or invariants (Hansen and Schmidt 2003). A more structured approach is to model mental states as multi-variable configurations (states) governed by constraints on how they evolve over sequences of state changes,

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

enabling verification of unwanted influence in their long-term evolution. Action languages (Gelfond and Lifschitz 1998), such as such as $\mathcal{A}$ (Gelfond and Lifschitz 1998), $\mathcal{C}$ (Gelfond and Lifschitz 1998), and $\mathcal{C}_{TAID}$ (Dworschak *et al.* 2008), provide a formal framework for describing and reasoning about actions and their effects on states of a domain. In action languages, actions are defined by their preconditions and effects. The preconditions specify the conditions that must hold true in the current state for an action to be applicable or executable. The high level declarative nature of action languages provide a suitable platform for eliciting and designing mental models and their principles of change, and characterizing them as dynamic systems.

In this paper, we introduce the action language $\mathcal{C}_{MT}$ (Mind Transition Language), designed to model changes in human mental states in response to observable actions. Through its translation into Answer Set Programming (ASP) (Brewka, Eiter and Truszczyński 2011), the language supports the generation and verification of action trajectories – sequences of actions and resulting states – while enforcing rules that forbid undesirable mental state transitions.

The proposed framework rests on two fundamental pillars: psychological theories of emotion and non-monotonic reasoning via ASP. Psychological theories offer structured frameworks to explain various aspects of human cognition, emotion, and behavior. Appraisal theories, such as the appraisal theory of emotion (Roseman 1996) which is modeled as an example in this work, identify classes of determinants and how combinations of them define emotions. From this, we can formalize fluents (changeable variables) and states (configurations of fluents) in our framework. Hence, in the proposed framework, a mental state of a human agent is approximated through a multi-dimensional representation, where each dimension corresponds to a factor in a psychological theory. In this way, we create abstractions of mental states, such as emotions, transforming psychological theories into computational models.

While action languages serve as intuitive specifications for dynamic reasoning processes, they can be characterized as logic programs by a mapping to ASP. The non-monotonic reasoning approach enabled by ASP complements these psychological foundations with a computational mechanism to handle the inherent variability in emotional reasoning. As a result, $\mathcal{C}_{MT}$ provides a logic programming foundation for knowledge elicitation and engineering of mental state dynamics. Emotional reasoning is inherently complex due to the exponential space of emotion states, particularly when accounting for their dynamics. This complexity poses a computational challenge, demanding solutions to combinatorial problems. ASP is well suited to address these challenges: it can express NP-search problems that are solvable by a nondeterministic Turing machine in polynomial time, with solutions encoded as answer sets (Brewka, Eiter and Truszczyński 2011). Thus, logic programming, and ASP in particular, offers an effective method for implementing and managing emotional reasoning. This further enables $\mathcal{C}_{MT}$ to capture dynamic properties such as *invariance* (Hansen and Schmidt 2003), understood in transition systems as a property that must hold in all states along every possible trajectory. In our setting, invariance specifies properties over mental state changes such that, when satisfied, the evolution of states follows a corresponding psychological principle of change, which can be rigorously analyzed through formal verification. Ultimately, the framework

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

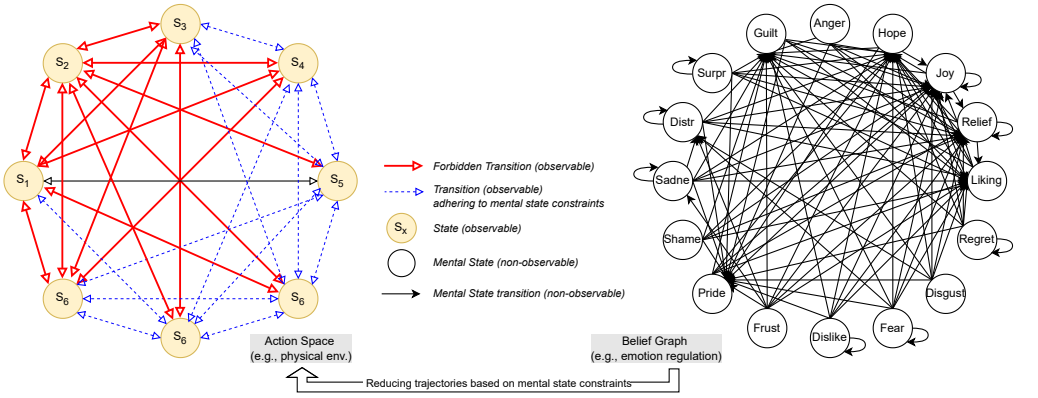


Fig 1. In $\mathcal{C}_{MT}$, available sequences of actions and states (trajectories) in the "physical" state space are constrained by the actions' influence on the "mental" state space.

supports reasoning about the evolution of human mental states while simultaneously constraining actions in the "physical" environment (see Figure 1).

From this basis, we present the following key contributions:

- We introduce the action language $\mathcal{C}_{MT}$, serving as a foundational platform for capturing properties relevant to mental state reasoning.
- We present the framework's syntax and semantics, and its practical implementation in ASP, evaluated formally and empirically.
- We present characterizations of $\mathcal{C}_{MT}$ in the setting of emotional reasoning, capturing psychological theories such as Appraisal theory of Emotion (Roseman 1996), Hedonic Emotion Regulation (Zaki 2020) and Utilitarian Emotion Regulation (Tamir *et al.* 2007).
- We present how $\mathcal{C}_{MT}$ can be applied as a method for representing, verifying and comparing psychological theories in terms of action trajectories.

A key distinction between our approach and traditional statistical methods lies in the deterministic nature of the latter versus the non-deterministic nature of ASP in our framework. Statistical methods typically collapse a decision into a single point, which can lead to a loss of information and a lack of flexibility in reasoning processes. In contrast, ASP's non-determinism yields multiple answer sets, offering a range of possible solutions that can be filtered and analyzed according to the introduced constraints and principles of change grounded in psychological theories. This approach not only preserves the richness of potential outcomes but also allows for a more nuanced and comprehensive understanding of the dynamics of human mental states.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 introduces the formal framework. Section 4 presents a case study in emotional reasoning. Section 5 provides a formal analysis of the framework. Section 6 reports on the experimental evaluation. Section 7 illustrates the approach through an example in human emotion verification. Section 8 offers a broader discussion of the results. Finally, Section 9 concludes the paper and outlines directions for future work.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

2 Related work

Human behavior has been formally analyzed at multiple levels and abstractions, ranging from integrative physiology (Fass and Lieber 2009) at a low level to formalizations of social practices (Erdogan *et al.* 2025) at a high level. On the level of mental states (Pereira *et al.* 2007; Jiang *et al.* 2007; Ong *et al.* 2019; Erdogan *et al.* 2025), research has modeled mental-state context, such as simulating emotional behavior (Ong *et al*. 2019) or modeling expected human responses to affective states (Jiang *et al.* 2007). Moreover, a diverse body of research has explored the dynamics of mental states (Rao and Georgeff 1995; Ramırez and Geffner 2009; Shvo and McIlraith 2020); Keren *et al.* 2014). Plan recognition as planning, introduced by Ramirez and Geffner (Ramırez and Geffner 2009), applies classical artificial intelligence (AI) planning techniques to infer the goals and plans of agents based on observed actions. Active Goal Recognition (Shvo and McIlraith 2020) extends goal recognition using contingent planning and landmark-based hypothesis elimination, enabling an active observer to sense, reason, and act. Furthermore, in Epistemic planning (Bolander and Andersen 2011), a generalization of classical planning, involves agents specifying goals that include the epistemic state, such as beliefs, of other agents. Recent extensions of epistemic planning to cognitive planning formalizes a method for influencing the cognitive state of the target agent (Davila *et al.* 2021; Lorini *et al.* 2022).

While modal logic underpins many approaches to reasoning about knowledge and belief, it typically represents mental states as sets of possible worlds, often reducing complex phenomena such as "emotions" to atomic propositions (e.g. "Joy" or "Sadness") rather than multi-variable configurations with structured interdependencies. This limits expressivity in capturing the causes, constraints, and consequences of mental-state transitions over time. To address this, various logics of mental attitudes and emotions (Adam *et al.* 2009; Lorini and Schwarzentruber 2011; Dastani and Lorini 2012; Steunebrink *et al.* 2012; Lorini 2021) have been developed. These works integrate epistemic logic with additional structures, such as plausibility orderings for belief strength (Lorini 2021), logic-based appraisal models for emotions (Adam *et al.* 2009), and STIT (Seeing-To-It-That) logic for counterfactual reasoning about emotions such as regret and rejoicing (Lorini and Schwarzentruber 2011). A notable approach is the TOMA framework (Erdogan *et al.* 2025), which integrates epistemic logic to model and update beliefs while introducing higher-order abstractions based on lower-order beliefs. By employing modal operators for belief and knowledge, the system facilitates structured reasoning about trust, social roles, and norms.

In the setting of mental state representation and reasoning, two related areas of research regard: (1) Logics of mental attitudes and emotion (Adam *et al.* 2009; Lorini and Schwarzentruber 2011; Dastani and Lorini 2012; Steunebrink *et al.* 2012; Lorini 2021), and (2) as previously mentioned, epistemic planning (Bolander and Andersen 2011) extended to cognitive planning (Davila *et al.* 2021; Lorini *et al.* 2022). Logics of mental attitudes and emotion aim to formalize the relationships between epistemic and motivational attitudes of human and artificial agents, as well as the influence of mental attitudes on emotions. Some related works in this line of research include the logical formalization of OCC theory of emotions (Adam *et al.* 2009), The formalization of

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

counterfactual emotions (Lorini and Schwarzentruber 2011), the representation of emotion intensity and coping strategies (Dastani and Lorini 2012), the modeling of emotion triggers (Steunebrink *et al.* 2012), and the logical theory of epistemic and motivational attitudes and their dynamics (Lorini 2021). Nevertheless, the principles constraining this causality and its potential side effects are not considered. In contrast to previous approaches to modeling and reasoning about mental states, the proposed $\mathcal{C}_{MT}$ action language deals with a multi-dimensional representation of mental states and the constraints for modeling principle-based transitions between them.

In the area of affective agents and computational theory of mind, agent models have been developed to reason about emotion and behavior (Si *et al.* 2010; Ong *et al.* 2019; Jara-Ettinger 2019; Yongsatianchot and Marsella 2021; Qiu *et al.* 2022). For instance, agents based on Partially Observable Markov Decision Processes (Yongsatianchot and Marsella 2021) have been used to model emotion, showing potential in simulating human behavior. Nevertheless, the state space is solely in the physical environment. Hence, they have lacked to capture mental-state dynamics to reason about causes for mental-states and mental transitions. Inverse Reinforcement Learning has been proposed to model Theory of Mind (Ong *et al.* 2019), for predicting peoples' actions and inferring mental states through policy reconstruction. However, a limitation with their approach is that it assumes identical, and rational, decision-making due to its inability to capture the variability in human reasoning, neglecting individual differences and emotional influences that may significantly impact human behavior. To effectively model the dynamics of the human mind, capturing individual and contextual factors, a low level, multi-dimensional approach is required. A study on socially-aware agents (Qiu *et al.* 2022) proposes a hybrid mental state parser that extracts information from dialogue and event observations to maintain a graphical representation of an agent's mind. The nodes represent agents, their personas, objects, and descriptions of the setting. The edges between these nodes depict the state of mind of the agents, capturing how their beliefs and mental states change as the setting evolves. While their representation can represent mental state changes, it lacks the ability to represent constraints of changes for analyzing their adherence to specific principles of mental change. Also, the graphical representation is on a high level, in contrast to the multidimensional representation proposed in the current paper.

Let us also mention that there is a range of approaches in the setting of BDI (Belief-Desire-Intention) agents (Pereira *et al.* 2007; Jiang *et al.* 2007; Jones *et al.* 2009; Sánchez-López and Cerezo 2019; Sánchez *et al.* 2019), integrating affective states into traditional BDI models, such as to simulate expected emotional responses (Pereira *et al.* 2007) or provide architectures for embedding emotion-driven reasoning (Jiang *et al.* 2007). While these works have considered affective states, such as emotion, integrated the BDI model, challenges persist in implementing affective constraints, such as emotion regulation (Sánchez-López and Cerezo 2019).

The related paradigm of Human-Aware Planning (Ahrndt *et al.* 2014) focuses on planning in the state-space of the physical environment, often in shared contexts where both autonomous agents and humans act. The planner typically reasons about the human's physical actions and coexists with them in domains such as robot navigation

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

or collaborative manipulation. However, for a rational software agent to anticipate how its actions or events may influence a human's mental state, it must reason in a space that goes beyond the physical. This requires a planning model that explicitly includes the state-space of the human mind: mental states, allowable transitions between them, and the actions that may trigger such transitions.

From this background, we highlight that there is a lack of formal treatment on principled constraints governing how mental states evolve over an interaction. While psychological theories, such as emotion regulation (Tamir *et al*. 2008; Ortner *et al*. 2018), describe patterns of mental-state change, formal models rarely integrate such principles, for example as transition constraints or properties of invariance (Hansen and Schmidt 2003), which can be rigorously evaluated using formal methods. By enforcing such constraints, a system can verify that mental-state trajectories, in contrast to isolated transitions, adhere to predefined principles while enabling the detection of deviations or violations. A more structured approach represents mental states as multi-variable configurations governed by trajectory-level constraints that regulate their long-term evolution. We refer to such a framework for defining valid mental states and transitions as a BG, formally introduced and analyzed in this paper.

In the following sections, we introduce the formal framework that integrates these concepts, providing a structured approach to modeling and managing the dynamics of human mental states.

3 Formal framework

This section introduces the syntax and semantics of the proposed formal framework and the action language extension $\mathcal{C}_{MT}$, which builds on ASP-based action reasoning. Although $\mathcal{C}_{MT}$ is implementation-agnostic and can, in principle, be realized using different computational frameworks, we view it as a high-level formal language designed specifically for realization in Answer Set Programming (ASP). From this perspective, we construct expressions in $\mathcal{C}_{MT}$ to align naturally with their ASP encodings. The framework offers flexibility to specialize for specific mental state domains, such as emotions, aligning with established psychological theories, such as the Appraisal theory of Emotion (Roseman 1996). By incorporating principles of mental change, sets of transition constraints are formalized and implemented in terms of integrity constraints in answer set programs.

In the proposed action language, similar to previous action languages, *fluents* represent properties that can change over time. These can describe various aspects of human-agent interactions, including non-observable aspects such as psychological attributes and observable aspects of agents or the environment. The value of a fluent at any given time depends on how it is affected by so-called *actions* or indirectly by other fluents.

We build on the action language $\mathcal{C}_{TAID}$ (Dworschak *et al.* 2008), originally designed for modeling biological systems. $\mathcal{C}_{TAID}$ includes features such as *allowance*, *inhibition*, and *triggers*, which are also relevant for modeling the dynamics of mental states. The *triggers* causal law accounts for interactions based on reactions, making it particularly useful for capturing how mental states may change as indirect effects of actions or external events. *Allowance* rules specify that an action can occur under certain conditions

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

but is not mandatory, while *inhibition* rules prevent an action from occurring in certain contexts. These mechanisms are especially important in modeling mental states, where dependencies between cognitive and emotional factors may be partially known or not explicitly modeled. In such cases, allowance and inhibition rules provide a flexible way to account for uncertainties and exceptions in mental state reasoning. These rules are relevant in the context of mental states where we have partial knowledge about the dependencies and reasons behind interactions in the mind. In situations where dependencies are partially known or not explicitly modeled, such as some conditions of the environment, allowance and inhibition rules provide a flexible way to handle uncertainties and exceptions.

We introduce the action language $\mathcal{C}_{MT}$ (Mind Transition Language), serving as a foundational platform, capturing domain-independent properties and constraints relevant to mental state reasoning. Along with the action language, we introduce some "syntactic sugar" to facilitate expressions about mental state dynamics. This includes "influences mental fluent" (similar to "causes" in $\mathcal{C}_{TAID}$), handling actions or events that may influence a change in a mental state, and the rules "facilitates" and "contravenes" (similar to "allowance" and "inhibition" rules in $\mathcal{C}_{TAID}$), which regulate actions' execution in an initial state. However, we specialize these rules to particularly concern *human actions* regulated by *mental fluents*. This is motivated by emotion theories, suggesting that "emotions have distinctive goals and action tendencies" (Roseman *et al*. 1994).

Moreover, $\mathcal{C}_{MT}$ introduces abstractions that we call **mental states**, which are defined by sets of fluents, along with mental state transition constraints, called **forbids to cause** rules, that specify relationships among mental fluents from one state to the next. Notably, such expressions enabling direct restriction of fluents between states, independent of actions, have not been explicitly incorporated in previous action languages like $\mathcal{C}_{TAID}$.

Consequently, the proposed framework for mental-state reasoning is comprised of two components: (1) The action language $\mathcal{C}_{MT}$, which defines actions in the environment/interaction that trigger changes in mental states, and (2) a set of constraints that precisely define valid transitions between mental states, derived from psychological principles. These constraints characterize a so-called BG. In ASP, a BG is encoded as sets of integrity constraints, restricting particular fluent changes in transitions between mental states. Mental state dynamics are linked to actions in the environment by considering their effects on fluents in the mental state abstractions. In this way, a BG filters the potential trajectories resulting from the action language, based on their effects on mental states (see Figure 2 for a conceptualization of the framework).

Intuitively, $\mathcal{C}_{MT}$ extends $\mathcal{C}_{TAID}$ as follows:

- To begin with, it refines the action and fluent alphabet by distinguishing between *observable* environmental fluents (denoted $\mathbf{F}^E$) and actions (denoted $\mathbf{A}^E$) and *non-observable* mental fluents (denoted $\mathbf{F}^H$) and human actions (denoted $\mathbf{A}^H$).
- It introduces specialized constructs such as `influences`, `facilitates`, and `contravenes`, which replicate $\mathcal{C}_{TAID}$'s `causes`, `allows`, and `inhibits` rules but are designed specifically to model how mental states regulate human behavior. This is made explicit in the ASP encoding, where these rules specifically are linked to mental

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

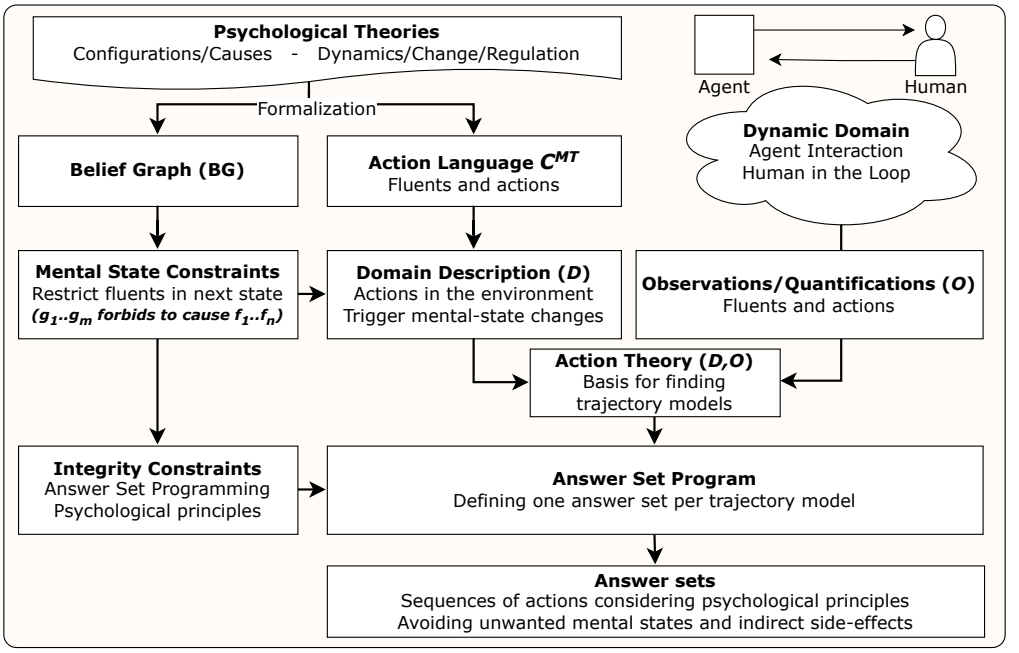


Fig 2. Conceptual framework.

fluents and human actions. While these constructs do not provide a substantial formal contribution, they support knowledge elicitation regarding modeling of mental dynamics, in line with one of the core purposes behind action languages.

- It incorporates a novel constraint rule,

$$(g_1^h, \ldots, g_m^h \textbf{ forbids to cause } f_1^h, \ldots, f_n^h)$$

specifying that if the mental fluents $g_1^h, \ldots, g_m^h \in \mathbf{F}^H$ hold at time t, then the mental fluents $f_1^h, \ldots, f_n^h \in \mathbf{F}^H$ are forbidden from holding at time $t+1$. This construct enables the specification of state transitions that are invalid regardless of any action occurrence – something not expressible in $\mathcal{C}_{TAID}$.

- A contribution of the language, utilizing the new **forbids to cause** rule, is the overall methodology for formalizing psychological theories in terms of dynamic computational models: abstractions of sets of fluents called mental-states and valid transitions between them. A set of these forbidding rules, on top of a set of dynamic causal laws, defines a so-called BG, which captures valid mental state transitions and filters action trajectories with constraints based on principles from psychological theories.

These additions enhance $\mathcal{C}_{MT}$ by enabling it to model complex dynamics of mental states and verify system behavior in accordance with psychological principles. The novel **forbids to cause** rule is essential in mental state modeling for two key reasons: (1) actions may directly induce undesirable changes in mental fluents, and (2) such changes may indirectly trigger unwanted ramification effects on other mental fluents. Consequently, this rule provides a formal mechanism to restrict actions that *directly* or *indirectly* lead to adverse mental outcomes.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

3.1 Belief graph

A BG is a set of propositional atoms for valid transitions between mental states. The mental states within a BG represent configurations of factors that contribute to specific mental states. These factors, known as mental fluents, are determined by psychological theories and encompass a range of possible values, thus defining the potential states within a BG. These mental fluents are defined as follows:

Definition 1 (Mental fluent)
Let $C = \{c_1, \ldots, c_n\}$ be a set of symbols denoting psychological classes, and let $V = \{V_{c_1}, \ldots, V_{c_n}\}$ be a set of total ordered sets of constants denoting psychological values for each class of C. A mental fluent is a ground atom $f(c, v)$ of arity 2 such that $c \in C$, $v \in V_c$.

A mental state space S is defined as the set of all possible combinations of mental fluents.

Definition 2 (Mental state space)
Let $C = \{c_1, \ldots, c_n\}$ be a set of psychological classes, and let $V = \{V_{c_1}, \ldots, V_{c_n}\}$ be a set where each $V_{c_i} (1 \leq i \leq n)$ is the set of possible values for the class c_i. The mental state space S is defined as:

$$S = \bigcup_{v_1 \in V_{c_1}, v_2 \in V_{c_2}, \ldots, v_n \in V_{c_n}} \{f(c_1, v_1), f(c_2, v_2), \ldots, f(c_n, v_n)\}$$

where each set $\{f(c_1, v_1), f(c_2, v_2), \ldots, f(c_n, v_n)\}$ denotes a unique combination of mental fluents, called a mental state.

We now proceed by defining a BG with states S and directed edges $E \subseteq S \times S$. These edges between states capture principles of mental change suggested by psychological theories, providing rules for allowed mental state transitions.

Definition 3 (Belief Graph)
A BG is a directed graph $BG = (S, E)$ where S is a finite set of mental states and $E \subseteq S \times S$ is a set of directed edges representing valid transitions between mental states.

A BG needs to be characterized for each specific application since different approaches to mental influence may be applicable. Consequently, to support controlled system behavior and principle-based assessment, it is essential to establish reasoning principles of "monotonicity" governing valid transitions between mental states. These principles can be motivated from various sources, such as psychological theories or insights from human experts. In the setting of reasoning about transitions between emotions, there are different principles and theories that have contrasting views on emotional change. For instance, principles of *hedonic emotion regulation* (Zaki 2020) aim to increase positive emotions and decrease negative emotions. Another set of principles regards *utilitarian emotion regulation* (Tamir *et al.* 2007), which aim to increase emotions that provide utility, such as control. Hence, the BG is agnostic to the choice of psychological theories. We show how this flexibility enables the framework to model, and compare, psychological theories through the principle-based assessment of generated trajectories.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

3.2 The $\mathcal{C}_{MT}$ action language

$\mathcal{C}_{MT}$ consists of a set of symbols representing actions and fluents, forming the alphabet of the action language. Given that the purpose of an action language is to provide a higher-level framework that makes modeling dynamic systems more natural and modular, several aspects of the language serve as syntactic sugar to simplify the modeling of mental states. The primary addition, beyond a specialized alphabet, is the incorporation of **forbids to cause** rules that given a set of mental fluents in the current state forbids a set of mental fluents to hold in the next state. By capturing elements of both the external environment and the human mind, the language facilitates the description and analysis of how environmental events influence human mental states and behavior, and through the new forbidding constraints, *actions* as well as *indirect effects* on mental fluents can be constrained. We require this addition to the language in order to model principles of mental change.

Definition 4 ($\mathcal{C}_{MT}$ 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 quantified non-observable features of mental-states of humans. $\mathbf{F}^E$ and $\mathbf{F}^H$ are pairwise disjoint.
- $\mathbf{A} = \mathbf{A}^E \cup \mathbf{A}^H$ 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.

Within $\mathcal{C}_{MT}$, a domain description defines static and dynamic causal laws for actions. These laws precisely express the expected influences exerted on mental fluents, either as direct effects of actions or as indirect causal effects. The laws governing mental change operate by modulating the variables of a given mental state while adhering to the constraints outlined by the BG.

Definition 5 ($\mathcal{C}_{MT}$ domain description language)
The $\mathcal{C}_{MT}$ domain description language $D^{MT}(\mathbf{A}, \mathbf{F})$ consists of static and dynamic causal laws of the following form:

$\mathcal{C}_{MT}$ domain description language (extending $\mathcal{C}_{TAID}$):

$(a$ **causes** $f_1, \ldots, f_n$ **if** $g_1, \ldots, g_m)$ (1)
$(f_1, \ldots, f_n$ **if** $g_1, \ldots, g_m)$ (2)
$(f_1, \ldots, f_n$ **triggers** $a)$ (3)
$(f_1, \ldots, f_n$ **allows** $a)$ (4)
$(f_1, \ldots, f_n$ **inhibits** $a)$ (5)
$($**noconcurrency** $a_1, \ldots, a_n)$ (6)
$($**default** $g)$ (7)

Mental-state domain description language extension:

$(a$ **influences** $f_1^h, \ldots f_n^h$ **if** $g_1, \ldots, g_m)$ (8)
$(g_1, \ldots, g_m$ **influences** $f_1^h, \ldots, f_n^h)$ (9)
$(g_1^h, \ldots, g_m^h$ **facilitates** $a^h)$ (10)
$(g_1^h, \ldots, g_m^h$ **contravenes** $a^h)$ (11)
$(g_1^h, \ldots, g_m^h$ **forbids to cause** $f_1^h, \ldots, f_n^h)$ (12)

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

where $a \in \mathbf{A}$, $a^h \in \mathbf{A}^H$, and $a_i \in \mathbf{A}$ $(0 \leq i \leq n)$ and $f_j \in \mathbf{F}$, $(0 \leq j \leq n)$ and $g_j \in \mathbf{F}$, $(0 \leq j \leq m)$, and $f_1^h, \ldots, f_n^h \in \mathbf{F}^H$, $g_1^h, \ldots, g_m^h \in \mathbf{F}^H$ are mental fluents of the form $f(c, v)$, where c is a psychological class and v is a psychological value.

We next provide the formal specification of the action language components in terms of states, rules, and transition conditions. Once this foundation is in place, we proceed to their characterization under answer set semantics, where the operational meaning of the laws is captured by logic programs and their answer sets.

BGs are expressed in terms of $\mathcal{C}_{MT}$ logic programs, that is finite sets of dynamic and causal laws on mental states. This characterization allows us to ensure controlled mental change by restricting the states and state-transitions. The mental state, within the context of the domain description, represents an interpretation of the current state of the system.

Definition 6 (Mental state interpretation)
A state $s \in S$ of the domain description $D^{MT}(\mathbf{A}, \mathbf{F})$ is an interpretation over $\mathbf{F}$ such that

1. for every static causal law $(f_1, \ldots, f_n$ **if** $g_1, \ldots g_m) \in D^{MT}(\mathbf{A}, \mathbf{F})$, we have $\{f_1, \ldots, f_n\} \subseteq s$ whenever $\{g_1, \ldots g_m\} \subseteq s$.
2. for every static causal law $(g_1, \ldots, g_m$ **influences** $f_1^h, \ldots, f_n^h) \in D^{MT}(\mathbf{A}, \mathbf{F})$, we have $\{f_1^h, \ldots, f_n^h\} \subseteq s$ whenever $\{g_1, \ldots, g_m\} \subseteq s$, and $\{f_1^h, \ldots, f_n^h\} \subseteq \mathbf{F}^H$.

S denotes all the possible states of $D^{MT}(\mathbf{A}, \mathbf{F})$.

In this definition, a state is determined by the satisfaction of static causal laws within the domain description $D^{MT}(\mathbf{A}, \mathbf{F})$. The first condition ensures that if the prerequisite mental fluents for an action are true, then the consequent mental fluents will also be true in the state. The second condition specifies that if certain mental fluents influence a particular mental fluent, then the influenced mental fluent must be true when all the influencing mental fluents are true.

Let us define the laws of the domain description more precisely.

Definition 7 (Domain description)
By considering the domain description $D^{MT}(\mathbf{A}, \mathbf{F})$ and a state s, the following rules and laws apply:

1. An inhibition rule $(f_1, \ldots, f_n$ **inhibits** a) is active in s, if $f_1, \ldots, f_n \in s$, otherwise, passive. The set $A_I(s)$ is the set of actions for which there exists at least one active inhibition rule in s (as in $\mathcal{C}_{TAID}$ (Dworschak *et al.* 2008)).
2. A triggering rule $(f_1, \ldots, f_n$ **triggers** a) is active in s, if $f_1, \ldots, f_n \in s$ 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 $\overline{A}_T(s)$ is the set of actions for which there exists at least one triggering rule and all triggering rules are passive in s (as in $\mathcal{C}_{TAID}$ (Dworschak *et al.* 2008)).
3. An allowance rule $(f_1, \ldots, f_n$ **allows** a) is active in s, if $f_1, \ldots, f_n \in s$ and all inhibition rules of action a are passive in s, otherwise, the allowance rule is passive

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

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 $\overline{A}_A(s)$ is the set of actions for which there exists at least one allowance rule and all allowance rules are passive in s (as in $\mathcal{C}_{TAID}$ (Dworschak *et al.* 2008)).

4. A facilitating rule ($g_1^h, \ldots, g_m^h$ **facilitates** a^h) is active in s, if $a^h \in \mathbf{A}^H$ and $g_1^h, \ldots, g_m^h \in s$ and all inhibition rules and contravening rules of action a are passive in s, otherwise, the facilitating rule is passive in s. The set $A_{FAC}(s)$ is the set of actions for which there exists at least one active facilitating rule in s. The set $\overline{A}_{FAC}(s)$ is the set of actions for which there exists at least one facilitating rule and all facilitating rules are passive in s.
5. An contravening rule ($g_1^h, \ldots, g_m^h$ **contravenes** a^h) is active in s, if $a^h \in \mathbf{A}^H$ and $g_1^h, \ldots, g_m^h \in s$ and all inhibition rules and facilitating rules of action a are passive in s, otherwise, the contravening rule is passive in s. The set $A_{INT}(s)$ is the set of actions for which there exists at least one active contravening rule in s.
6. A dynamic causal law (a causes $f_1, \ldots, f_n$ if $g_1, \ldots, g_m$) is applicable in s, if $g_1, \ldots, g_m \in s$.
7. A static causal law ($f_1, \ldots, f_n$ if $g_1, \ldots, g_m$) is applicable in s, if $g_1, \ldots, g_m \in s$.
8. A dynamic causal law (a **influences** $f_1^h, \ldots, f_n^h$ if $g_1, \ldots, g_m$) is applicable in s, if $g_1, \ldots, g_m \in s$, and $f_1^h, \ldots, f_n^h \in F^H$, and $f_i \in F (1 \leq i \leq n)$.
9. A static causal law ($g_1, \ldots, g_m$ **influences** $f_1^h, \ldots, f_n^h$) is applicable in s, if $g_1, \ldots, g_m \in s$, and $f_1^h, \ldots, f_n^h \in F^H$, and $f_i \in F (1 \leq i \leq n)$.
10. A forbidding rule ($g_1^h, \ldots, g_m^h$ **forbids to cause** $f_1^h, \ldots, f_n^h$) is active in s if $\{g_1^h, \ldots, g_m^h\} \subseteq s$, where $f_1^h, \ldots, f_n^h, g_1^h, \ldots, g_m^h \in \mathbf{F}^H$. The set $F(s)$ is the set of fluents that are forbidden in $s+1$ when at least one active forbidding rule exists in s.

Intuitively, the rules describe how actions may or may not occur, and how fluents are connected across and within states. Rule (1) states that certain fluents can inhibit an action, making it inactive in the current state. Rule (2) captures when fluents trigger an action, meaning it must occur if its triggers are satisfied and it is not inhibited. Rule (3) specifies when fluents allow an action, meaning the action is permitted but not enforced. Rules (4) and (5) extend this to human actions: some mental fluents can facilitate a human action, while others can contravene it. Rule (6) introduces dynamic causal laws, describing how actions under certain conditions cause fluents to hold in the next state. Rule (7) introduces static causal laws, which specify that if some fluents hold in a state, then others must co-hold in the same state. Rule (8) adds dynamic influence for mental fluents: when an action occurs under certain conditions, it brings about specified mental fluents in the next state. Rule (9) specifies static influence for mental fluents: when its conditions hold, the corresponding mental fluents must also hold in the same state. Finally, Rule (10) introduces forbids to cause, which regulates mental change by specifying that if certain mental fluents hold in the current state, then some mental fluents are not allowed to appear in the next state.

The output of the action language is in terms of trajectories. A mental-state trajectory consists of a sequence of valid transitions, represented as

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

$\langle s_0, A_1, s_1, A_2, \ldots, A_n, s_n \rangle$, with alternating sets of mind-altering actions $A \subseteq \mathbf{A}$ and mental-states $s \in S$, following the constraints of the BG.

Definition 8 (Trajectory)
Let $D^{MT}(\mathbf{A, F})$ be a domain description. A trajectory $\langle s_0, A_1, s_1, A_2, \ldots, A_n, s_n \rangle$ of $D^{MT}(\mathbf{A, F})$ is a sequence of sets of actions $A_i \subseteq A$ and states s_i of $D^{MT}(\mathbf{A, F})$ satisfying the following conditions for $0 \leq \mathrm{i} < \mathrm{n}$:

1. $(s_i, A, s_{i+1}) \in S \times 2^A \backslash \{\} \times S$
2. $A_T(s_i) \subseteq A_{i+1}$
3. $A_{FAC}(s_i) \subseteq A_{i+1}$
4. $\overline{\mathrm{A}}_T(s_i) \cap A_{i+1} = \emptyset$
5. $\overline{\mathrm{A}}_A(s_i) \cap A_{i+1} = \emptyset$
6. $A_I(s_i) \cap A_{i+1} = \emptyset$
7. $\overline{\mathrm{A}}_{FAC}(s_i) \cap A_{i+1} = \emptyset$
8. $A_{INT}(s_i) \cap A_{i+1} = \emptyset$
9. $|A_i \cap B| \leq 1 \; for \; all \; (noconcurrency B) \in D^{MT}(\mathbf{A, F})$.
10. $F(s_i) \cap s_{i+1} = \emptyset$ (no forbidden fluents in s_{i+1})

The set B in condition (9) represents a subset of actions restricted by noconcurrency constraints in $D^{MT}(\mathbf{A}, \mathbf{F})$, ensuring that actions in B cannot execute simultaneously. This prevents conflicts where multiple actions attempt to modify the same fluent at the same time. While actions affecting distinct fluents can occur concurrently, those altering the same fluent are mutually exclusive, maintaining consistency in fluent updates and ensuring well-defined state transitions.

Definition 9 (Action Observation Language)
The action observation language of $\mathcal{C}_{MT}$ (similar to $\mathcal{C}_{TAID}$) consists of expressions of the following form:

$$(f \textbf{ at } t) \qquad (a \textbf{ occurs_at } t) \qquad (8)$$

where $f \in \mathbf{F}$, a is an action and $t \in \mathbb{N}$ is a point in time.

The action observation language allows us to specify observations concerning the current state of mental fluents and the execution of actions that influence mental states. By combining observations with the causal laws of the domain description, we can generate plans, explanations, and predictions regarding the behavior of the system. This integration of observations and the domain description is referred to as an action theory.

Definition 10 (Action Theory)
Let D be a domain description and O be a set of observations. The pair (D, O) is called an action theory.

The action theory forms the basis for constructing trajectory models, trajectories where all observations are satisfied, providing a structured representation of the system's dynamics over time. Trajectory models enable us to analyze and reason about the evolution of mental states and actions, allowing for a deeper understanding how the mental state domain operates and how it responds to different observations and changes.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Definition 11 (Trajectory Model)
Let (D, O) be an action theory. A trajectory $\langle s_0, A_1, s_1, A_2, \ldots, A_n, s_n\rangle$ of D is a trajectory model of (D, O), if it satisfies all observations of O in the following way:

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

We can observe that actions in a trajectory model can be actions executed by a rational software agent to influence mental fluents, or actions estimated to be executed by a human agent.

An important consequence of the **forbids to cause** rule is that it can render an action theory inconsistent. Since such a rule blocks any transition to a state containing its forbidden fluents, a trajectory that reaches such a state is invalid. If all candidate trajectories are thus blocked, then no trajectory model exists and the action theory is inconsistent. This regulative effect is intentional, ensuring that unwanted mental fluents cannot appear in valid states.

Definition 12 (Action Theory Consistency)
An action theory (D, O) is consistent iff there exists a trajectory model of (D, O). Otherwise, (D, O) is inconsistent.

The central computational task is to determine, for a given $\mathcal{C}_{MT}$ action theory (D, O), whether there exists a trajectory model of (D, O). This baseline feasibility check underpins subsequent reasoning tasks such as planning, explanation, and query answering.

Definition 13 ($\mathcal{C}_{MT}$ Decision Problem)
Given a $\mathcal{C}_{MT}$ theory (D, O), decide if there exists a trajectory model of (D, O).

We need a mechanism to write queries about state dynamics. The Action Query Language provides a means to inquire about specific sequences of actions and their impact on fluents and states. This is achieved by specifying subsets of the action set and their corresponding occurrences in time.

Definition 14 (Action Query Language)
The action query language of $\mathcal{C}_{MT}$ regards assertions about executing sequences of actions with expressions that constitute trajectories. A query is of the following form: $(f_1, \ldots, f_n$ **after** A_i **occurs_at** $t_i, \ldots, A_m$ **occurs_at** $t_m)$ where $f_1, \ldots, f_n$ are fluent literals $\in \mathbf{F}$, $A_i, \ldots, A_m$ are subsets of $\mathbf{A}$, and $t_i, \ldots, t_m$ are points in time.

By formulating queries in the action language, causal relationships between actions and their effects can be investigated, contributing to a deeper understanding of system dynamics, informed decision-making and controlled methods for automated planning.

Definition 15 (Query Truth in Action Theory)
Let (D, O) be an action theory and Q a query.

- Q holds skeptically in (D, O) iff $\tau \models Q$ for every trajectory model τ of (D, O).
- Q holds credulously in (D, O) iff $\tau \models Q$ for some trajectory model τ of (D, O).

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Given a $\mathcal{C}_{MT}$ action theory (D, O) and a query Q, the decision problem regards whether Q holds in all (or some) trajectory models of (D, O).

3.3 Action language in answer set semantics

In this section, we provide operational semantics for $\mathcal{C}_{MT}$ by translation into answer set programs. The goal is to construct an encoding such that the encoding has exactly one answer set for every trajectory model of a $\mathcal{C}_{MT}$ theory. Intuitively, each answer set represents a trajectory model: for a trajectory $\langle s_0, A_1, s_1, \ldots, A_n, s_n \rangle$, the atoms `holds`$(f, t)$ indicate that fluent f holds in state s_t, and the atoms `holds(occurs`(a),t) indicate that action a occurs between s_t and s_{t+1}. Conversely, each answer set uniquely determines a trajectory by the fluents and actions it contains, thereby establishing a one-to-one correspondence between answer sets and trajectory models and reducing the central reasoning problems of $\mathcal{C}_{MT}$ to ASP computation.

To make the presentation self-contained, we include translations from (Dworschak *et al.* 2008) for the $\mathcal{C}_{TAID}$ language (expressions 1–7 in Definition 5), and then extend them with the new constructs of $\mathcal{C}_{MT}$. We organize the translations into distinct subsections, addressing the Action Description Language, the Action Observation Language, and the Action Query Language.

3.3.1 Encoding of the action description language

We consider the encoding of $\mathcal{C}_{TAID}$ action description language, which we extend with components of $\mathcal{C}_{MT}$.

Define symbols for a fluent f $\in \mathbf{F^E}$ and an action a $\in \mathbf{A^E}$ (extending (Dworschak *et al.* 2008)).

```
fluent_e(f), action_e(a).
fluent(f) :- fluent_e(f).
action(a) :- action_e(a).
```

Define symbols for a fluent e $\in \mathbf{F^H}$ and an action u $\in \mathbf{A^H}$ (extending (Dworschak *et al.* 2008)).

```
mental_fluent(e), human_action(u).
fluent(e) :- mental_fluent(e).
action(u) :- human_action(u).
```

Define a range of time points $0 \leq t \leq t_max$, $t \in \mathbb{N}$ (as in (Dworschak *et al.* 2008)).

```
time(0..t_max). #const t_max = N.
```

Contradiction constraint (as in (Dworschak *et al.* 2008)): A fluent $f \in \mathbf{F}$ and its negation $\neg f$ cannot hold simultaneously at the same time step $T \in \mathbb{N}$. In ASP, this is enforced as a constraint, ensuring that `holds`(f, T) and `holds(neg(f),T)` cannot both be true. The predicates `fluent`(f) and `time`(T) declare fluents and time steps.

```
:- holds(f,T), holds(neg(f),T), fluent(f), time(T).
```

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Inertial fluents (as in (Dworschak *et al.* 2008)): An inertial fluent persists across time steps unless modified by an action or a static causal law. In ASP, this is encoded by ensuring that if `holds`(f, T) is true and not overridden, then `holds`$(f, T+1)$ holds by default. The predicates `fluent`(f) and `time`(T) track fluents over time.

```
holds(f,T+1) :- holds(f,T), not holds(neg(f,T+1)), not default(f),
fluent(f), time(T), time(T+1).
```

Non-inertial fluents (as in (Dworschak *et al.* 2008)): A non-inertial fluent resets to a default value unless explicitly updated. In ASP, this is encoded by inferring `holds`(f, T) whenever `default`(f) holds and `holds(neg(f),T)` is not inferred. The predicates `fluent`(f), `default`(f), and `time`(T) define non-inertial fluents and their default behavior.

```
holds(f,T) :- not holds(neg(f),T), default(f), fluent(f), time(T).
```

Dynamic causal law (as in (Dworschak *et al.* 2008)): A dynamic causal law (a **causes** f **if** $g_1, \ldots, g_n$) states that if an action $a \in \mathbf{A}$ occurs at time $T \in \mathbb{N}$ and fluents $g_1, \ldots, g_n \in \mathbf{F}$ hold, then a fluent $f \in \mathbf{F}$ must hold at time $T+1$. In ASP, one rule is generated for each dynamic causal law, ensuring that if `holds(occurs`(a, T)`)` and `holds`$(g_1, T), \ldots,$ `holds`(g_n, T) hold, then `holds`$(f, T+1)$ is inferred. The predicates `fluent`(f) and `fluent`$(g_1), \ldots,$ `fluent`(g_n) declare fluents, `action`(a) declares the action, and `time`(T) ensures valid time progression.

```
holds(f,T+1) :- holds(occurs(a),T),
holds(g_1,T), ..., holds(g_n,T),
fluent(g_1), ..., fluent(g_n),
fluent(f), action(a), time(T), time(T+1).
```

Static causal law (as in (Dworschak *et al.* 2008)): A static causal law (f **if** $g_1, \ldots, g_n$) states that a fluent $f \in \mathbf{F}$ holds at time $T \in \mathbb{N}$ if fluents $g_1, \ldots, g_n \in \mathbf{F}$ hold at the same time. In ASP, one rule is generated for each static causal law, ensuring that if `holds`$(g_1, T), \ldots,$ `holds`(g_n, T) hold, then `holds`(f, T) is inferred. The predicates `fluent`(f) and `fluent`$(g_1), \ldots,$ `fluent`(g_n) declare fluents, and `time`(T) ensures valid time progression.

```
holds(f,T) :- holds(g_1,T), ..., holds(g_n,T),
fluent(g_1), ..., fluent(g_n), fluent(f), time(T).
```

Inhibition rule (as in (Dworschak *et al.* 2008)): An inhibition rule ($f_1, \ldots, f_n$ **inhibits** a) states that an action $a \in \mathbf{A}$ is prevented from occurring at time $T \in \mathbb{N}$ if fluents $f_1, \ldots, f_n \in \mathbf{F}$ hold. In ASP, one rule is generated for each inhibition rule, ensuring that if `holds`$(f_1, T), \ldots,$ `holds`(f_n, T) hold, then `holds(ab(occurs`$(a), T)$`))` is inferred, marking the action as inhibited. The predicates `fluent`(f_1), ..., `fluent`(f_n) declare fluents, `action`(a) declares the action, and `time`(T) ensures valid time progression.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

```
holds(ab(occurs(a)),T) :- holds(f_1,T), ..., holds(f_n,T),
action(a), fluent(f_1), ..., fluent(f_n), time(T).
```

As in $\mathcal{C}_{TAID}$, unlike standard ASP planning encodings, we do not use a choice rule for action generation, such as
{ holds(occurs(A), T): action(A) } = 1 :- T=1.t_max. Instead, actions are determined by logical conditions: they occur if they are explicitly triggered, allowed, and not inhibited. This ensures that only valid and necessary actions are selected, avoiding unnecessary non-determinism.

Triggering rule (as in (Dworschak *et al.* 2008)): A triggering rule ($f_1, \ldots, f_n$ **triggers** a) states that an action $a \in \mathbf{A}$ occurs at time $T \in \mathbb{N}$ if fluents $f_1, \ldots, f_n \in \mathbf{F}$ hold and no inhibition rule is active. In ASP, one rule is generated for each triggering rule, ensuring that if $\texttt{holds}(f_1, T), \ldots, \texttt{holds}(f_n, T)$ hold and $\texttt{not holds(ab(occurs}(a, T)))$, then $\texttt{holds(occurs}(a, T))$ is inferred. The predicates $\texttt{fluent}(f_1), \ldots, \texttt{fluent}(f_n)$ declare fluents, $\texttt{action}(a)$ declares the action, and $\texttt{time}(T)$ ensures valid time progression.

```
holds(occurs(a),T) :- not holds(ab(occurs(a)),T),
holds(f_1,T), ..., holds(f_n,T),
fluent(f_1), ..., fluent(f_n), action(a), time(T).
```

Allowance rule (as in (Dworschak *et al.* 2008)): An allowance rule ($f_1, \ldots, f_n$ **allows** a) states that an action $a \in \mathbf{A}$ is permitted to occur at time $T \in \mathbb{N}$ if fluents $f_1, \ldots, f_n \in \mathbf{F}$ hold and no inhibition rule is active. In ASP, one rule is generated for each allowance rule, ensuring that if $\texttt{holds}(f_1, T), \ldots, \texttt{holds}(f_n, T)$ hold and $\texttt{not holds(ab(occurs}(a, T)))$, then $\texttt{holds(allow(occurs}(a), T))$ is inferred. The predicates $\texttt{fluent}(f_1), \ldots, \texttt{fluent}(f_n)$ declare fluents, $\texttt{action}(a)$ declares the action, and $\texttt{time}(T)$ ensures valid time progression.

```
holds(allow(occurs(a)),T) :- not holds(ab(occurs(a)),T),
holds(f_1,T), ..., holds(f_n,T), fluent(f_1), ..., fluent(f_n),
action(a), time(T).
```

Ensure exogenous actions can always occur (as in (Dworschak *et al.* 2008)): This rule ensures that an action $a \in \mathbf{A}$ is always permitted to occur at any time step $T \in \mathbb{N}$, regardless of specific preconditions. In ASP, this is encoded by inferring $\texttt{holds(allow(occurs}(a), T))$ for every declared action $\texttt{action}(a)$ and time point $\texttt{time}(T)$. This guarantees that exogenous actions remain available throughout the execution.

```
holds(allow(occurs(a)),T) :- action(a), time(T).
```

No-concurrency constraint (as in (Dworschak *et al.* 2008)): The actions $a_1, \ldots, a_n \in \mathbf{A}$ cannot occur simultaneously at time $T \in \mathbb{N}$. In ASP, this is encoded as a constraint that ensures selecting two or more actions at T leads to inconsistency. The rule applies over declared actions $\texttt{action}(a_1), \ldots, \texttt{action}(a_n)$ and time points $\texttt{time}(T)$.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

```
:- time(T), 2 {holds(occurs(a_1),T) :
action(a_1), ..., holds(occurs(a_n),T) : action(a_n)}.
```

3.3.2 Encoding of the action observation language

We consider the encoding of $\mathcal{C}_{TAID}$ action observation language, which we extend with components of $\mathcal{C}_{MT}$. The action observation language allows for the specification of observations about fluents and actions over time. Observations provide constraints on the initial state and subsequent state transitions, ensuring that execution traces align with known facts.

Initial state fluent observations (as in (Dworschak *et al.* 2008)): A fluent $f \in \mathbf{F}$ that holds in the initial state at time $T = 0$ is directly asserted. In ASP, this is represented by $\texttt{holds}(f, 0)$, ensuring that observed fluents are set at the start of execution, forming the basis for reasoning about subsequent state transitions.

```
holds(f,0).
```

Fluent observations for other states (as in (Dworschak *et al.* 2008)): If a fluent $f \in \mathbf{F}$ is observed to hold at time $T \in \mathbb{N}$, it must be enforced. In ASP, this is encoded as a constraint ensuring that $\texttt{holds}(f, T)$ must be true whenever f is observed. The rule applies over declared fluents $\texttt{fluent}(f)$ and time points $\texttt{time}(T)$.

```
:- not holds(f,T), fluent(f), time(T).
```

Generate possible completions of the initial state (as in (Dworschak *et al.* 2008)): A fluent $f \in \mathbf{F}$ in the initial state at $T = 0$ must either hold or its negation $\neg f$ must hold, but not both. In ASP, this is encoded using a default completion principle, ensuring that $\texttt{holds}(f, 0)$ is inferred unless `holds(neg(f),0)` is explicitly stated, and vice versa. The rule applies over declared fluents $\texttt{fluent}(f)$ at time step $\texttt{time}(0)$.

```
holds(f,0) :- not holds(neg(f),0).
holds(neg(f),0) :- not holds(f,0).
```

Exogenous action observations (as in (Dworschak *et al.* 2008)): An exogenous action $a \in \mathbf{A}$ that occurs at time $T \in \mathbb{N}$ is recorded as a fact. This corresponds to an external execution of a, independent of triggering, allowance, or inhibition conditions. In ASP, this is represented by asserting $\texttt{holds}(\texttt{occurs}(a), T)$, ensuring that the action is registered as having taken place. The rule applies over declared actions $\texttt{action}(a)$ and time points $\texttt{time}(T)$.

```
holds(occurs(a),T).
```

Observed non-occurrence of actions (as in (Dworschak *et al.* 2008)): If an action $a \in \mathbf{A}$ does not occur at time $T \in \mathbb{N}$, this must be explicitly recorded. This corresponds to cases where no triggering or allowance rule is active, or an inhibition rule prevents execution. In ASP, this is enforced as a constraint, ensuring that $\texttt{holds}(\texttt{neg}(\texttt{occurs}(a), T))$ is

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

inferred when the action does not occur. The rule applies over declared actions `action`(a) and time points `time`(T).

```
:- not holds(neg(occurs(a)),T), action(a), time(T).
```

Action execution conditions (as in (Dworschak *et al.* 2008)): An action $a \in \mathbf{A}$ occurs at time $T \in \mathbb{N}$ if at least one allowance rule ($f_1, \ldots, f_n$**allows** a) is active, no inhibition rule ($f_1, \ldots, f_n$**inhibits** a) applies, and no external constraint negates its occurrence. In ASP, this is encoded by ensuring that `holds(occurs`$(a), T)$ is inferred under these conditions. Additionally, if an action does not occur, its negation `holds(neg(occurs`$(a), T))$ is inferred. The rule applies over declared actions `action`(a) and time points `time`(T), where $T < t_{\max}$.

```
holds(occurs(a),T) :- holds(allow(occurs(a)),T),
not holds(ab(occurs(a)),T), not holds(neg(occurs(a)),T),
action(a), time(T), T < t_max.

holds(neg(occurs(a)),T) :- not holds(occurs(a),T),
action(a), time(T), T < t_max.
```

3.3.3 Encoding of the action query language

We consider the encoding of $\mathcal{C}_{TAID}$ action query language, which we extend with components of $\mathcal{C}_{MT}$.

Goal achievement constraint (as in (Dworschak *et al.* 2008)): The goal must be achieved in every valid plan. This corresponds to ensuring that at least one fluent configuration satisfying the goal holds. In ASP, this is enforced as a constraint ensuring that `achieved` is inferred.

```
:- not achieved.
```

Initial goal satisfaction (as in (Dworschak *et al.* 2008)): The goal is considered achieved if it already holds at the initial time step $T = 0$.

```
achieved :- achieved(0).
```

Persistence of goal achievement (as in (Dworschak *et al.* 2008)): If the goal is achieved at time $T + 1$ but was not achieved at T, then it remains achieved for all subsequent time steps. The predicates `time`(T) ensure valid time progression.

```
achieved :- achieved(T+1), not achieved(T), time(T), time(T+1).
```

Fluent-based goal satisfaction (as in (Dworschak *et al.* 2008)): The goal is achieved at time T if a set of fluents $f_1, \ldots, f_n \in \mathbf{F}$ required for goal satisfaction hold. This corresponds to a static causal law of the form ($f_1, \ldots, f_n$**if** goal_conditions). In ASP, this is encoded by inferring `achieved`(T) when `holds`$(f_1, T), \ldots,$ `holds`(f_n, T) hold.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

```
achieved(T) :- holds(f_1, T), ..., holds(f_n, T),
achieved(T+1), fluent(f_1), ..., fluent(f_n), time(T), time(T+1).
```

Final goal satisfaction (as in (Dworschak *et al.* 2008)): The goal must hold at the maximum time step $t_{\max}$. This ensures that the trajectory satisfies the required final state.

```
achieved(t_max) :- holds(f_1, t_max), ..., holds(f_n, t_max),
fluent(f_1), ..., fluent(f_n).
```

Execution of allowed actions (as in (Dworschak *et al.* 2008)): An action $a \in \mathbf{A}$ occurs at time $T \in \mathbb{N}$ if it is allowed by an active allowance rule ($f_1, \ldots, f_n$**allows** a), is not inhibited by an inhibition rule ($f_1, \ldots, f_n$**inhibits**a), and the goal has not yet been achieved. In ASP, this ensures that `holds(occurs`$(a), T$`)` is inferred only when these conditions hold. The predicates `fluent`$(f_1), \ldots,$ `fluent`(f_n) declare fluents, `action`(a) declares actions, and `time`(T) defines time steps.

```
holds(occurs(a), T) :- holds(allow(occurs(a)), T), not achieved(T),
not holds(ab(occurs(a)), T), not holds(neg(occurs(a)), T),
action(a), time(T).
```

Explicit non-occurrence of actions (as in (Dworschak *et al.* 2008)): If an action $a \in \mathbf{A}$ does not occur at time $T \in \mathbb{N}$, this must be explicitly recorded. This corresponds to enforcing that an action not chosen in the trajectory is negated.

```
holds(neg(occurs(a)), T) :- not holds(occurs(a), T),
action(a), time(T).
```

3.3.4 Encoding of the mental state language extension

We now proceed by presenting the translations for static and dynamic causal laws for mental state specifications introduced in $\mathcal{C}_{MT}$.

Dynamic causal law for mental fluents (introduced in $\mathcal{C}_{MT}$): A dynamic causal law (a **influences** $f_1^h, \ldots, f_n^h$ **if** $g_1, \ldots, g_m$) specifies that an action $a \in \mathbf{A}$ causes mental fluents $f_1^h, \ldots, f_n^h \in \mathbf{F}^H$ to hold at $T+1$ if fluents $g_1, \ldots, g_m \in \mathbf{F}$ hold at $T \in \mathbb{N}$. The ASP encoding generates a rule for each mental fluent $f_i^h \in \mathbf{F}^H$ $(1 \leq i \leq n)$, of the form:

```
holds(f_i, T+1) :- holds(occurs(a), T),
    holds(g_1, T), ..., holds(g_m, T),
    fluent(g_1), ..., fluent(g_m), mental_fluent(f_i),
    action(a), time(T).
```

Static causal law for mental fluents (introduced in $\mathcal{C}_{MT}$): A static causal law ($g_1, \ldots, g_m$ **influences** $f_1^h, \ldots, f_n^h$) states that if fluents $g_1, \ldots, g_m \in \mathbf{F}$ hold at time T, then mental fluents $f_1^h, \ldots, f_n^h \in \mathbf{F}^H$ also hold at $T \in \mathbb{N}$. The ASP encoding generates a rule for each mental fluent $f_i^h \in \mathbf{F}^H$ $(1 \leq i \leq n)$, of the form:

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

```
holds(f_i, T) :- holds(g_1, T), ..., holds(g_m, T),
        fluent(g_1), ..., fluent(g_m), mental_fluent(f_i),
        time(T).
```

The following translations regard contravening and facilitating rules (introduced in $\mathcal{C}_{MT}$). Unlike the "inhibition" and "allowance" rules outlined in $\mathcal{C}_{TAID}$ (Dworschak *et al.* 2008), which impact actions in the current time step, "contravenes" and "facilitates" relate particularly about a human action $a^h \in \mathbf{A}^H$ and mental fluents $f^h \in \mathbf{F}^H$. As we have previously discussed; appraisal theories have suggested that emotional states have distinctive "action tendencies" (Roseman *et al.*1994). Capturing these semantics, while seemingly redundant due to the general rules in the action language, supports knowledge representation.

Facilitation rule (introduced in $\mathcal{C}_{MT}$): A facilitation rule ($g_1^h, \ldots, g_m^h$ **facilitates** a^h) states that if mental fluents $g_1^h, \ldots, g_m^h \in \mathbf{F}^H$ hold at time $T \in \mathbb{N}$, and the action $a^h \in \mathbf{A}^H$ is not inhibited, then a^h occurs. The ASP encoding ensures that `holds(occurs`$(a^h), T$`)` is inferred if `not holds(ab(occurs`$(a^h), T$`))` holds.

```
holds(occurs(a), T) :- not holds(ab(occurs(a)), T),
        holds(g_1, T), ..., holds(g_m, T),
        mental_fluent(g_1), ..., mental_fluent(g_m),
        human_action(a), time(T).
```

Contravening rule (introduced in $\mathcal{C}_{MT}$): A contravening rule ($g_1^h, \ldots, g_m^h$ **contravenes** a^h) states that if mental fluents $g_1^h, \ldots, g_m^h \in \mathbf{F}^H$ hold at time $T \in \mathbb{N}$, then the human action $a^h \in \mathbf{A}^H$ is inhibited from occurring. The ASP encoding infers `holds(ab(occurs`$(a^h), T$`))` under these conditions.

```
holds(ab(occurs(a)), T) :- holds(g_1, T), ..., holds(g_m, T),
        mental_fluent(g_1), ..., mental_fluent(g_m),
        human_action(a), time(T).
```

Forbidding rule (introduced in $\mathcal{C}_{MT}$): A forbidding rule ($g_1^h, \ldots, g_m^h$ **forbids to cause** $f_1^h, \ldots, f_n^h$) enforces constraints on mental state transitions in the BG. If mental fluents $g_1^h, \ldots, g_m^h \in \mathbf{F}^H$ hold at time $T \in \mathbb{N}$, then fluents $f_1^h, \ldots, f_n^h \in \mathbf{F}^H$ are forbidden to hold at $T+1$. This ensures that the belief graph's constraints on state transitions are respected. The ASP encoding introduces an integrity constraint to prohibit transitions that violate these restrictions.

```
:- holds(f_1, T+1), ..., holds(f_n, T+1),
        holds(g_1, T), ..., holds(g_m, T),
        mental_fluent(f_1), ..., mental_fluent(f_n),
        mental_fluent(g_1), ..., mental_fluent(g_m),
        time(T).
```

In the upcoming section, we will explore a case study that applies a specialization of $\mathcal{C}_{MT}$ for emotional reasoning. The case study models emotion states and dynamics based

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

on well-established psychological theories. By examining this specialized application, we aim to demonstrate the effectiveness and versatility of the $\mathcal{C}_{MT}$ framework for modeling different kinds of mental state dynamics.

4 Case study: Emotional reasoning

The section presents a characterization of emotion dynamics in $\mathcal{C}_{MT}$ with specialized subsets of fluents and mental state constraints adapted for emotional reasoning. This is achieved by formalizing different emotion theories; the Appraisal theory of Emotion (AE) (Roseman 1996), Hedonic Emotion Regulation (HER) (Zaki 2020), and Utilitarian Emotion Regulation (UER) (Tamir *et al.* 2007), capturing links between human emotions and their underlying causes in the environment. To this end, subsets of mental state fluents, called emotion fluents, together with a set of action rules and sets of mental state constraints are specified, capturing principles from emotion theories. By following the psychological theory of AE, an emotion state-space is defined with 108 emotional configurations (states), through which a set of 16 basic human emotions, according to AE, can be represented and explained. By following the theories of HER and UER, different constraints for emotional-change are defined. HER focuses on augmenting positive emotions while diminishing negative ones. Conversely, UER seeks to promote particular emotions, including potentially negative ones, that enhance specific attributes, such as motivation or control, serving a utilitarian purpose in the long run. Consequently, HER and UER adopt contrasting principles for effecting emotional change. By examining the trajectories produced through the application of either HER or UER within the state-space of AE, we can compare and evaluate their respective behaviors.

4.1 Emotion theories: AE, HER and UER

The **Appraisal theory of emotion (AE)** (Roseman 1996) proposes that emotions are caused by an appraisal of a situation in terms of 1) being consistent or inconsistent with needs, importance of the situation, 2) being consistent or inconsistent with goals, the attainability/potential to achieve goals, 3) who/what is accountable/caused the situation, which can be the environment, others, or oneself, and 4) as being easy or difficult to control. According to AE, the difference between goal consistency/attainability and need consistency/importance determines negative, stable and positive emotions. More intense negative emotions (e.g. Anger or Fear) arise when the need consistency is greater than the goal consistency, while less intense negative emotions can arise when both the need consistency and goal consistency are low. On the other hand, positive emotions (e.g. Joy or Liking) arise when the goal consistency is greater than the need consistency, or when both are high. By ranking consistency values as $Low < Undecided < High$ and by looking at the difference between need and goal consistency, positive and negative emotions can be distinguished.

Hedonic emotion regulation (HER) (Zaki 2020) is a theory for regulating emotions, guided by the goals to increase positive emotion and decrease negative emotion. According to HER, both of these emotion regulation goals are associated with improved well-being, where decreasing of negative emotion has been most effective (Ortner *et al.*

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Table 1. *Hedonic emotion regulation as basis for principles of change*

Hedonic Emotion Regulation	Principle of Change
Need Consistency/Importance: A high need consistency is associated with positive emotions only when the goal consistency is high, but negative if the goal consistency is low (supported by (Bühler *et al.* 2019)).	High need consistency maximizes positive emotion only when goal consistency is high.
Goal Consistency/Attainability: A high goal consistency is associated with positive emotions. In particular when the need consistency is high (supported by (Bühler *et al.* 2019)).	High goal consistency maximizes positive emotion.
Accountability: Accountability does not affect whether the emotion is negative or positive. Nevertheless, when the environment is perceived as primarily accountable for the situation, less negative emotions appear than when an individual (self/other) is perceived as accountable (supported by (Passyn and Sujan 2006)).	Accountability to environment maximizes positive emotions.
Control Potential: Control Potential does not affect whether the emotion is negative or positive. Nevertheless, a control potential above low is often associated with accommodating emotions, closer to positive emotions (supported by (Roseman 1996)).	High control potential maximizes positive emotion when there is a non-negative balance between goal consistency and need consistency, that is, when goal consistency is equal or higher than need consistency.

2018). The principles of HER can be applied to reason about emotional change. For instance, the relation between goal attainability and goal importance has been empirically explored, showing that goal attainability, rather than goal importance, was positively linked to well-being (Bühler *et al.* 2019). Another empirical study analyzed self-responsibility and emotions, and showed that accountability does not by itself affect whether the emotion is negative or positive. Nevertheless, when something else than when an individual (self/other) is perceived as accountable, less negative emotions appear (Passyn and Sujan 2006). Another empirical study examined the relationship between control potential and emotions (Roseman 1996). The study found that the perception of high or low control potential influenced the experience of accommodating emotions or contending emotions, respectively. However, the study did not find a direct effect of control potential on the experience of positive or negative emotions.

By analyzing the four dimensions of emotion as proposed by AE (Need consistency, Goal consistency, Accountability, and Control potential), we can derive meaningful interpretations about the relationship between HER and AE in the context of hedonic emotional change. These interpretations are summarized in Table 1.

UER (Tamir *et al.* 2007) is a theory emotion regulation that, in contrast to HER, is guided by the goals to experience emotions in the short-term for a long-term utilitarian purpose. This may involve accepting a temporary negative emotion to increase their capability of dealing with a situation. For instance, valuing long-term goals (i.e. focusing

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

on need, importance and motivation) (Dweck 2017) over immediate pleasure makes individuals more likely to engage in UER. Furthermore, a study on goal-setting (Schunk 2001) found that when individuals perceive a discrepancy between their present performance and a desired goal, it can lead to dissatisfaction. This dissatisfaction, in turn, can serve as a motivator for increased effort and striving towards the goal. Another study, exploring accountability and emotion (Autry and Langenbach 1985), suggests that self-accountability for a situation has utilitarian gains by having effects on the individual's "locus of control." Moreover, a study exploring control and emotion (Tamir *et al.* 2008) found that increasing control potential, such as through an increased level of anger, can have utilitarian gains. Another study on the emotion of fear (Tamir and Ford 2009) proposes the concept of "fear as function for goal pursuit," highlighting its role in motivating goal-directed behavior. Yet another study (Parrott 2002) suggests that "negative" emotions can modify an individual's "readiness to think and act, and their potential functional utility."

By analyzing the four dimensions of emotion as proposed by AE (Need consistency, Goal consistency, Accountability, and Control potential), we can derive meaningful interpretations about the relationship between UER and AE in the context of utilitarian emotional change. These interpretations are summarized in Table 2.

Following the interpretations of HER and UER, we can define transition constraints in an emotion specialized BG. This will allow certain sets of trajectories depending on which emotional-change theory we model. In both cases, we use the state space defined by AE.

In the following section, we present a formalization of these theories in terms of $\mathcal{C}_{MT}$. We first define an emotion state in terms of the dimensions of AE. We then define a set of transition constraints that restrict particular configurations (states) of AE appraisals to arise. Two sets of transition constraints are defined, constituting two different subgraphs of valid transitions in the AE state-space, a so-called *hedonic emotion state* and an *utilitarian emotion state*, based on HER and UER principles, respectively.

4.2 Formalizing emotional reasoning

Components of AE and HER (resp. UER) are formalized as a specialized BG to reason about emotion states and hedonic (resp. utilitarian) emotional change to reduce unintended emotional side-effects. The constraints of the specialized BG serve as safety restrictions for emotion-influencing actions, applicable for different types of emotional reasoning settings.

Recall that AE defines emotions as a composition of an individual's appraisal of a situation, in terms of need consistency (importance), goal consistency (attainability), accountability (who/what) and control potential. By following this definition of emotional causes, we can define *emotion fluent*, a changeable emotion variable.

Definition 16 (Emotion fluent)
An emotion fluent is a mental fluent $f(c, v)$, a ground atom of arity 2, where $c \in C$, and $v \in V_c$, such that $C = \{ne, go, ac, co\}$ is a set of constants denoting psychological classes and $V = \{V_{ne}, V_{go}, V_{ac}, V_{co}\}$ a set of total ordered sets of constants denoting psychological values for each class of C, where:

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Table 2. *Utilitarian emotion regulation as basis for principles of change*

Utilitarian Emotion Regulation	Principle of Change
Need consistency/importance: Valuing long-term goals over immediate pleasure makes individuals more likely to engage in UER. They are willing to endure temporary negative emotions to maximize utility and achieve important goals (supported by (Dweck 2017)).	High need consistency maximizes utility.
Goal consistency/attainability: Perceiving important goals as difficult to achieve in the short-term can motivate individuals to engage in UER strategies to increase their long-term success (supported by (Schunk 2001)).	Low or Undecided goal consistency maximizes utility if the need consistency is high.
Accountability: Individuals perceiving themselves as primarily accountable for their long-term goals drive them to regulate emotions in a utilitarian manner. Taking responsibility for emotional experiences enhances their ability to deal with challenges effectively. Furthermore, if the control potential is high, the environment can be perceived as accountable, where an individual may be able to affect aspects of the environment for utilitarian gains (supported by (Autry and Langenbach 1985)).	Accountability to self maximizes utility. Accountability to environment maximizes utility if the control potential is high.
Control potential: Sense of control is an utilitarian characteristic. Accepting and regulating temporary negative emotions to increase perceived control contribute to long-term goal attainment (supported by (Tamir *et al.* 2008)).	High control potential maximizes utility.

$$V_{ne} = \{high, low, undecided\}$$
$$V_{go} = \{high, low, undecided\}$$
$$V_{ac} = \{self, other, environment, undecided\}$$
$$V_{co} = \{high, low, undecided\}.$$

where *ne* denotes need consistency, *go* denotes goal consistency, *ac* denotes accountability and *co* denotes control potential.

By defining a set of emotions following AE in this way, and by utilizing different principles of emotion regulation, we can specify preferred (e.g. hedonic or utilitarian) transitions between emotion states. In the following subsection, we specify a BG to reason about emotional transitions.

Following the AE-theory, 16 emotion states are specified, corresponding to each basic emotion: {Anger, Dislike, Disgust, Sadness, Hope, Frustration, Fear, Distress, Joy, Liking, Pride, Surprise, Relief, Regret, Shame, Guilt}. It is important to note that various other

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Anger			
ne:h	go:l	ac:o	co:h

Hope			
ne:u	go:h	ac:e	co:l

Joy			
ne:h	go:h	ac:e	co:u

Relief			
ne:h	go:h	ac:e	co:u

Dislike			
ne:u	go:l	ac:o	co:l

Frustration			
ne:h	go:l	ac:e	co:h

Liking			
ne:u	go:h	ac:o	co:u

Regret			
ne:u	go:l	ac:s	co:l

Disgust			
ne:l	go:l	ac:e	co:h

Fear			
ne:u	go:l	ac:e	co:l

Pride			
ne:u	go:h	ac:s	co:u

Shame			
ne:l	go:l	ac:s	co:h

Sadness			
ne:h	go:l	ac:e	co:l

Distress			
ne:l	go:l	ac:e	co:l

Surprise			
ne:u	go:u	ac:e	co:u

Guilt			
ne:h	go:h	ac:s	co:h

Fig 3. Emotion states following the appraisal theory of emotion (Roseman 1996). Each emotion state is here expressed by an intuitive "emotion" label on top, and an appraisal configuration consisting of a set of variable: value pairs below. The variables are: ne = need consistency, go = goal consistency, ac = accountable, co = control potential. The values are: l = low, h = high, u = undecided, o = other, s = self, e = environment.

emotion states can be defined (in total 108) by different combinations of the four AE dimensions, and it is not necessary, nor always preferred, to assign a specific label, such as "Joy," to each of them. While labels aid in expressing these states in a human readable way, they do not inherently contribute to the functionality of a system's reasoning. It is worth noting that emotional expressions vary across individuals, cultures, languages, and other factors. Therefore, it is crucial for the system to analyze emotion states using a multidimensional format. We can model the states and transitions as an EG that represents valid emotional change given a recognized emotion state configuration. In Figure 3, examples of emotion states are presented.

The specialized belief graph for emotion is captured by a given program specified by the semantics of the action language $\mathcal{C}_{MT}$, with constraints in terms of HER or UER, serving as restrictions for hedonic contra utilitarian emotional change, presented in the following section.

4.3 Action language emotion specification

The $\mathcal{C}_{MT}$ alphabet is specialized with an emotion related vocabulary and causal laws to specify fluents and actions according to AE-theory.

The emotional reasoning semantics is characterized by the constraints of a specialized BG for emotion, specified through a set of static causal laws and mental state constraints. In this way, we can restrict states and state-transitions to comply with principles of emotional change.

For any particular application, we need to define a BG that, based on application specific interaction goals and relevant theories, avoids unintended mental states. This specifies an BG with a subset of transitions (in the fully connected graph) that is considered valid. We define a hedonic theory specification that follows principles of HER, aiming to increase positive emotion and decrease negative emotion (see more details of HER in Section 4.1). In an abstraction of HER, the hedonic emotion specification is defined to not allow entering a negative emotion state and to preserve a positive balance in emotion.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Recall Figure 1, which illustrates the BG representing possible trajectories over environmental and mental states. Transitions (edges) in this graph are defined by a set of dynamic causal laws of the form (a **causes** $f_1, \ldots, f_n$ **if** $g_1, \ldots, g_m$), capturing how actions produce observable changes. Overlaid on these are forbidding rules of the form ($g_1^h, \ldots, g_m^h$ **forbids to cause** $f_1^h, \ldots, f_n^h$), which specify mental state transitions that are disallowed based on psychological theory. The red arrows in the graph correspond to environmental transitions that are excluded because they would cause such forbidden mental transitions. In this way, the Belief Graph integrates both causal dynamics and psychological constraints to filter out inadmissible action trajectories.

The hedonic theory specification can thus be formally expressed as follows.

Definition 17 (Hedonic emotion theory specification)
Let $D_{AE}^{MT}(\mathbf{A}, \mathbf{F})$ be a domain description over the $\mathcal{C}_{MT}$ action language, with a vocabulary consisting of emotion fluents (as defined in Definition 16) and actions over them. A hedonic theory specification in $D_{AE}^{MT}(\mathbf{A}, \mathbf{F})$ is a collection of causal laws as follows:

1. $f(c_{\text{ne}}, \text{high})$ **forbids to cause** $f(c_{\text{go}}, \text{low})$.
2. $f(c_{\text{ne}}, \text{high})$ **forbids to cause** $f(c_{\text{go}}, \text{undecided})$.
3. $f(c_{\text{ne}}, \text{high})$ **forbids to cause** $f(c_{\text{go}}, \text{high})$.
4. $f(c_{\text{ne}}, \text{undecided})$ **forbids to cause** $f(c_{\text{go}}, \text{low})$.
5. $f(c_{\text{ne}}, \text{undecided})$ **forbids to cause** $f(c_{\text{go}}, \text{undecided})$.
6. $f(c_{\text{ne}}, \text{undecided})$ **forbids to cause** $f(c_{\text{go}}, \text{high})$.
7. $f(c_{\text{go}}, \text{high})$ **forbids to cause** $f(c_{\text{go}}, \text{low})$.
8. $f(c_{\text{go}}, \text{high})$ **forbids to cause** $f(c_{\text{go}}, \text{undecided})$.
9. $f(c_{\text{go}}, \text{undecided})$ **forbids to cause** $f(c_{\text{go}}, \text{low})$.
10. $f(c_{\text{go}}, \text{low})$ **forbids to cause** $f(c_{\text{co}}, \text{high})$.
11. $\{f(c_{\text{ne}}, \text{high}), f(c_{\text{go}}, \text{low}), f(c_{\text{ac}}, \text{other})\}$ **forbids to cause** $f(c_{\text{co}}, \text{high})$.
12. $\{f(c_{\text{ne}}, \text{high}), f(c_{\text{go}}, \text{undecided}), f(c_{\text{ac}}, \text{other})\}$ **forbids to cause** $f(c_{\text{co}}, \text{high})$.
13. $\{f(c_{\text{ne}}, \text{high}), f(c_{\text{go}}, \text{low}), f(c_{\text{ac}}, \text{self})\}$ **forbids to cause** $f(c_{\text{co}}, \text{high})$.
14. $\{f(c_{\text{ne}}, \text{high}), f(c_{\text{go}}, \text{undecided}), f(c_{\text{ac}}, \text{self})\}$ **forbids to cause** $f(c_{\text{co}}, \text{high})$.
15. $\{f(c_{\text{ne}}, \text{high}), f(c_{\text{go}}, \text{low}), f(c_{\text{ac}}, \text{undecided})\}$ **forbids to cause** $f(c_{\text{co}}, \text{high})$.
16. $\{f(c_{\text{ne}}, \text{high}), f(c_{\text{go}}, \text{undecided}), f(c_{\text{ac}}, \text{undecided})\}$ **forbids to cause** $f(c_{\text{co}}, \text{high})$.

We further define an utilitarian emotion theory specification that follows principles of UER, aiming to prioritize particular emotional dimensions (e.g. need_consistency) that are associated with increased utilitarian gain (see more details of UER in Section 4.1). In an abstraction of UER, the utilitarian emotion theory specification is defined to allow entering a negative emotion state in the short-term if it may result in a long-term utilitarian gain.

Definition 18 (Utilitarian emotion theory specification)
Let $D_{AE}^{MT}(\mathbf{A}, \mathbf{F})$ be a domain description over the $\mathcal{C}_{MT}$ action language, with a vocabulary consisting of emotion fluents (as defined in Definition 16) and actions over them. A utilitarian theory specification in $D_{AE}^{MT}(\mathbf{A}, \mathbf{F})$ is a collection of causal laws as follows:

1. $f(c_{\text{ne}}, \text{low})$ **forbids to cause** $f(c_{\text{ne}}, \text{undecided})$.
2. $f(c_{\text{ne}}, \text{low})$ **forbids to cause** $f(c_{\text{ne}}, \text{high})$.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

3. $f(c_{\text{ne}}, \text{undecided})$ **forbids to cause** $f(c_{\text{ne}}, \text{low})$.
4. $f(c_{\text{ne}}, \text{undecided})$ **forbids to cause** $f(c_{\text{ne}}, \text{high})$.
5. $f(c_{\text{go}}, \text{low})$ **forbids to cause** $f(c_{\text{ne}}, \text{low})$.
6. $f(c_{\text{go}}, \text{low})$ **forbids to cause** $f(c_{\text{ne}}, \text{undecided})$.
7. $f(c_{\text{go}}, \text{undecided})$ **forbids to cause** $f(c_{\text{ne}}, \text{low})$.
8. $f(c_{\text{go}}, \text{undecided})$ **forbids to cause** $f(c_{\text{ne}}, \text{undecided})$.
9. $f(c_{\text{go}}, \text{high})$ **forbids to cause** $f(c_{\text{ne}}, \text{low})$.
10. $f(c_{\text{go}}, \text{high})$ **forbids to cause** $f(c_{\text{ne}}, \text{undecided})$.
11. $f(c_{\text{go}}, \text{high})$ **forbids to cause** $f(c_{\text{ne}}, \text{high})$.
12. $f(c_{\text{ac}}, \text{undecided})$ **forbids to cause** $f(c_{\text{ac}}, \text{self})$.
13. $f(c_{\text{ac}}, \text{undecided})$ **forbids to cause** $f(c_{\text{ac}}, \text{other})$.
14. $f(c_{\text{ac}}, \text{undecided})$ **forbids to cause** $f(c_{\text{ac}}, \text{environment})$.
15. $f(c_{\text{ac}}, \text{other})$ **forbids to cause** $f(c_{\text{ac}}, \text{undecided})$.
16. $f(c_{\text{ac}}, \text{other})$ **forbids to cause** $f(c_{\text{ac}}, \text{self})$.
17. $f(c_{\text{ac}}, \text{other})$ **forbids to cause** $f(c_{\text{ac}}, \text{environment})$.
18. $f(c_{\text{ac}}, \text{environment})$ **forbids to cause** $f(c_{\text{co}}, \text{low})$.
19. $f(c_{\text{ac}}, \text{environment})$ **forbids to cause** $f(c_{\text{co}}, \text{undecided})$.
20. $f(c_{\text{co}}, \text{low})$ **forbids to cause** $f(c_{\text{co}}, \text{undecided})$.
21. $f(c_{\text{co}}, \text{low})$ **forbids to cause** $f(c_{\text{co}}, \text{high})$.
22. $f(c_{\text{co}}, \text{undecided})$ **forbids to cause** $f(c_{\text{co}}, \text{low})$.
23. $f(c_{\text{co}}, \text{undecided})$ **forbids to cause** $f(c_{\text{co}}, \text{high})$.

These sets of **forbids to cause** rules are encoded as integrity constraints. Given a logic program P_{EG} (based on the translation in Section 3.3.1), we attach sets of integrity constraints representing HER (Listing 1) or UER (Listing 2) based principles. These integrity constraints are independent of the main logic program P_{EG}. By extending the program with different sets of integrity constraints, based on different emotion-regulation principles, we can analyze and compare their respective trajectories.

```
% HEDONIC EMOTION REGULATION
% Need consistency can only increase if goal consistency is high
:- holds(mental_fluent(need_consistency, high), T+1),
   holds(mental_fluent(goal_consistency, low), T), time(T).
:- holds(mental_fluent(need_consistency, high), T+1),
   holds(mental_fluent(goal_consistency, undecided), T), time(T).
:- holds(mental_fluent(need_consistency, undecided), T+1),
   holds(mental_fluent(goal_consistency, low), T), time(T).
:- holds(mental_fluent(need_consistency, undecided), T+1),
   holds(mental_fluent(goal_consistency, undecided), T), time(T).

% Goal consistency cannot decrease
:- holds(mental_fluent(goal_consistency, low), T+1),
   holds(mental_fluent(goal_consistency, V1), T), V1 != low, time(T).
:- holds(mental_fluent(goal_consistency, undecided), T+1),
   holds(mental_fluent(goal_consistency, V1), T), V1 != undecided, time(T).
```

Listing 1. Integrity Constraints for Hedonic Emotion Regulation.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

```
% Prevent high control potential if goal consistency is low
:- holds(mental_fluent(control_potential, high), T+1),
   holds(mental_fluent(goal_consistency, low), T), time(T).

% Prevent control potential from increasing to high
% when accountability is not the environment
% and there is a negative balance between goal and need consistency
:- holds(mental_fluent(control_potential, high), T+1),
   holds(mental_fluent(need_consistency, high), T),
   holds(mental_fluent(goal_consistency, low), T),
   holds(mental_fluent(accountability, other), T), time(T).

:- holds(mental_fluent(control_potential, high), T+1),
   holds(mental_fluent(need_consistency, high), T),
   holds(mental_fluent(goal_consistency, undecided), T),
   holds(mental_fluent(accountability, other), T), time(T).

:- holds(mental_fluent(control_potential, high), T+1),
   holds(mental_fluent(need_consistency, high), T),
   holds(mental_fluent(goal_consistency, low), T),
   holds(mental_fluent(accountability, self), T), time(T).

:- holds(mental_fluent(control_potential, high), T+1),
   holds(mental_fluent(need_consistency, high), T),
   holds(mental_fluent(goal_consistency, undecided), T),
   holds(mental_fluent(accountability, self), T), time(T).

:- holds(mental_fluent(control_potential, high), T+1),
   holds(mental_fluent(need_consistency, high), T),
   holds(mental_fluent(goal_consistency, low), T),
   holds(mental_fluent(accountability, undecided), T), time(T).

:- holds(mental_fluent(control_potential, high), T+1),
   holds(mental_fluent(need_consistency, high), T),
   holds(mental_fluent(goal_consistency, undecided), T),
   holds(mental_fluent(accountability, undecided), T), time(T).
```

Listing 1. Continued.

```
% UTILITARIAN EMOTION REGULATION
% Need consistency can only be influenced to high.
:- holds(mental_fluent(need_consistency, low), T+1),
    holds(mental_fluent(need_consistency, V1), T), V1 != low, time(T).
:- holds(mental_fluent(need_consistency, undecided), T+1),
    holds(mental_fluent(need_consistency, V1), T), V1 != undecided, time(T).

% The goal consistency can only be influenced to low or undecided,
% and the need consistency must be high.
:- holds(mental_fluent(goal_consistency, low), T+1),
    holds(mental_fluent(need_consistency, low), T), time(T).
:- holds(mental_fluent(goal_consistency, low), T+1),
    holds(mental_fluent(need_consistency, undecided), T), time(T).
```

Listing 2. Integrity Constraints for Utilitarian Emotion Regulation.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

```
:- holds(mental_fluent(goal_consistency, undecided), T+1),
    holds(mental_fluent(need_consistency, low), T), time(T).
:- holds(mental_fluent(goal_consistency, undecided), T+1),
    holds(mental_fluent(need_consistency, undecided), T), time(T).
:- holds(mental_fluent(goal_consistency, high), T+1),
    holds(mental_fluent(need_consistency, undecided), T), time(T).
:- holds(mental_fluent(goal_consistency, high), T+1),
    holds(mental_fluent(need_consistency, high), T), time(T).

% Accountability can only be influenced to self or environment.
% If environment, then the control must be high.
:- holds(mental_fluent(accountability, undecided), T+1),
    holds(mental_fluent(accountability, V1), T), V1 != undecided, time(T).
:- holds(mental_fluent(accountability, other), T+1),
    holds(mental_fluent(accountability, V1), T), V1 != other, time(T).
:- holds(mental_fluent(accountability, environment), T+1),
    holds(mental_fluent(control_potential, low), T), time(T).
:- holds(mental_fluent(accountability, environment), T+1),
    holds(mental_fluent(control_potential, undecided), T), time(T).

% Control potential can only be influenced to High.
:- holds(mental_fluent(control_potential, low), T+1),
    holds(mental_fluent(control_potential, V1), T), V1 != low, time(T).
:- holds(mental_fluent(control_potential, undecided), T+1),
    holds(mental_fluent(control_potential, V1), T), V1 != undecided, time(T).
```

Listing 2. Continued.

5 Formal analysis

We start by establishing a link between $\mathcal{C}_{MT}$ and answer set semantics, supporting implementations of the action language in ASP. In the proceeding parts of the section, we show that **forbids to cause** rules imply relations between mental fluents, allowing us to analyze Safety properties based on these relations (considering the invariance principle (Hansen and Schmidt 2003)) to be preserved in trajectories.

To support the formal analysis, we define an operator that transforms an action theory (and query, if applicable) into a logic program P_{MT} based on the high-level action language $\mathcal{C}_{MT}$.

Definition 19 ($\mathcal{C}_{MT}$ Logic Program)
Let D^{MT} be a $\mathcal{C}_{MT}$ domain description and O_{initial} a set of initial observations. We define the mapping function $P_{MT} := \Pi(D^{MT}, O_{\text{initial}})$ such that P_{MT} is a ground ASP program where, for every expression φ in $D^{MT} \cup O_{\text{initial}}$, $\Pi(\varphi)$ denotes the ASP translation of φ, and P_{MT} includes all such translations. Let $AS(P_{MT})$ denote the set of answer sets of the logic program P_{MT}.

In the following subsections, we proceed with the formal analysis.

5.1 Analysis of translation to answer set semantics

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

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Theorem 1
Let $(D^{MT}, O_{\text{initial}})$ *be an action theory such that* O_{initial} *specifies fluent observations in the initial state. Let* Q *be a query, and define:*

$$A_Q = \{(a \,\text{occurs_at}\, t_k) \mid a \in A_{k+1},\ 0 \leq k < t_{\max}\}.$$

Let $\mathcal{P}_{MT} := \Pi(D^{MT}, O_{\text{initial}} \cup A_Q)$ *be the* $\mathcal{C}_{MT}$ *logic program obtained via the translation operator* Π.

1. *If there is a trajectory model* $\langle s_0, A_1, s_1, \ldots, A_{t_{\max}}, s_{t_{\max}}\rangle$ *of* $\mathcal{C}_{MT}(D^{MT}, O_{\text{initial}} \cup A_Q)$, *then there is an answer set* $\mathcal{A}$ *of* $\mathcal{P}_{MT}$ *such that for all time points* $0 \leq k \leq t_{\max}$:
   (a) $holds(f, k) \in \mathcal{A}$ iff $f \in s_k$,
   (b) $holds(neg(f), k) \in \mathcal{A}$ iff $f \notin s_k$,
   (c) $holds(occurs(a), k) \in \mathcal{A}$ iff $a \in A_{k+1}$,
   (d) $holds(neg(occurs(a)), k) \in \mathcal{A}$ iff $a \notin A_{k+1}$.
2. *If there is an answer set* $\mathcal{A}$ *of* $\mathcal{P}_{MT}$ *and for each* $0 \leq k \leq t_{\max}$,
   (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 $\langle s_0, A_1, s_1, \ldots, A_{t_{\max}}, s_{t_{\max}}\rangle$ *of* $\mathcal{C}_{MT}(D^{MT}, O_{\text{initial}} \cup A_Q)$.

Proof
We prove Theorem 1 in two directions using induction over the maximal time point $t_{\max}$ of the trajectory. The proof applies the Splitting Set Theorem (Lifschitz and Turner 1994) to justify that answer sets can be constructed incrementally over time.

Part 1: From trajectory to answer set.
Let $\langle s_0, A_1, s_1, \ldots, A_{t_{\max}}, s_{t_{\max}}\rangle$ be a trajectory model of $\mathcal{C}_{MT}(D^{MT}, O_{\text{initial}} \cup A_Q)$. We define a set X of atoms such that:

$$\begin{aligned} X := &\{holds(f, k) \mid f \in s_k,\ 0 \leq k \leq t_{\max}\} \\ &\cup \{holds(neg(f), k) \mid f \notin s_k,\ 0 \leq k \leq t_{\max}\} \\ &\cup \{holds(occurs(a), k) \mid a \in A_{k+1},\ 0 \leq k < t_{\max}\} \\ &\cup \{holds(neg(occurs(a)), k) \mid a \notin A_{k+1},\ 0 \leq k < t_{\max}\}. \end{aligned}$$

Let P_k be the set of ASP rules in $\mathcal{P}_{MT}$ whose time arguments are $\leq k$. Then $\langle P_0, P_1, \ldots, P_{t_{\max}}\rangle$ forms a splitting sequence of $\mathcal{P}_{MT}$.
We prove by induction on k that $X_k := X \cap \text{Atoms}(P_k)$ is an answer set of P_k.

Base Case ($k = 0$). By construction of X_0, for every $f \in s_0$, we have $holds(f, 0) \in X_0$, and for every $f \notin s_0$, we have $holds(neg(f), 0) \in X_0$. These atoms match the translation of initial observations in O_{initial}. Hence, all rules in P_0 are satisfied, and X_0 is an answer set of P_0.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Induction Step. Assume X_k is an answer set of P_k. Consider P_{k+1}. The new rules introduced in $P_{k+1} \setminus P_k$ have head atoms with time $k+1$ and body atoms with times $\leq k+1$.

By the construction of X, and since $f \in s_{k+1}$ can only hold by one of the rules of Definition 8, we distinguish the following cases:

- **Case 1:** $f \in s_{k+1}$ due to inertia: $holds(f,k) \in X_k$, $neg(holds(neg(f),k+1)) \in X_{k+1}$.
- **Case 2:** $f \in s_{k+1}$ due to a dynamic causal law. There exists $a \in A_k$ and preconditions $g_1,\ldots,g_m \in s_k$ such that $holds(occurs(a),k)$, $holds(g_i,k) \in X_k$. The rule head $holds(f,k+1)$ is in X_{k+1}.
- **Case 3:** $f \in s_{k+1}$ due to a static causal law. Preconditions $g_1,\ldots,g_m \in s_{k+1}$ imply $holds(g_i,k+1) \in X_{k+1}$. Hence $holds(f,k+1) \in X_{k+1}$.
- **Case 4:** $f \in s_{k+1}$ by default (non-inertial fluent), that is no rule for $\neg f$ is applicable. Then $holds(neg(f),k+1) \notin X$, so $holds(f,k+1)$ is derived by the default rule.
- **Case 5:** $f \in s_{k+1}$ is mental: derived by static causal law as in Case 3.
- **Case 6:** $f^h \in s_{k+1}$ due to dynamic causal law. Same as Case 2 but with f^h, and $holds(f^h,k+1) \in X$.
- **Case 7:** $f^h \in s_{k+1}$ due to static causal law. As in Case 3.
- **Case 8:** $f^h \notin s_{k+1}$ due to a forbidding constraint $(g_1^h,\ldots,g_m^h$ **forbids to cause** $f^h) \in D^{MT}$ with $g_1^h,\ldots,g_m^h \in s_k$. The translated integrity constraint

$$:- \; holds(f^h,k+1),\; holds(g_1^h,k),\; \ldots,\; holds(g_m^h,k)$$

must be satisfied in X, so $holds(f^h,k+1) \notin X$.

Thus, all new rule bodies are satisfied, and all forbidden transitions are avoided. By the Splitting Set Theorem (Lifschitz and Turner 1994), X_{k+1} is an answer set of P_{k+1}. Therefore, X is an answer set of $\mathcal{P}_{MT}$.

Part 2: From answer set to trajectory.

Assume there is an answer set $\mathcal{A}$ of the logic program $\mathcal{P}_{MT}$. We construct a trajectory model

$$\langle s_0, A_1, s_1, \ldots, A_{t_{\max}}, s_{t_{\max}} \rangle$$

of $\mathcal{C}_{MT}(D^{MT}, O_{\text{initial}} \cup A_Q)$ as follows.

State and Action Set Construction. For every $0 \leq k \leq t_{\max}$, define:

$$s_k := \{f \mid holds(f,k) \in \mathcal{A}\} \cup \{\neg f \mid holds(neg(f),k) \in \mathcal{A}\},$$

$$A_{k+1} := \{a \mid holds(occurs(a),k) \in \mathcal{A}\}.$$

We will prove by induction on $t_{\max}$ that the constructed sequence satisfies all trajectory conditions of Definition 8.

Base Case ($t_{\max} = 0$): Then the trajectory reduces to $\langle s_0 \rangle$. The initial observations O_{initial} specify fluents and mental fluents that are true or false at time 0.

By translation of O_{initial} into facts of the form $holds(f,0)$ and $holds(neg(f),0)$ in $\mathcal{P}_{MT}$, and since $\mathcal{A}$ is an answer set, the state s_0 includes exactly the correct fluents and mental fluents. Hence, s_0 satisfies the initial conditions.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Induction Step: Assume that for $t_{\max} = r$, the sequence

$$\langle s_0, A_1, s_1, \ldots, A_r, s_r \rangle$$

is a trajectory model of $\mathcal{C}_{MT}$. We now extend it to

$$\langle s_0, A_1, s_1, \ldots, A_r, s_r, A_{r+1}, s_{r+1} \rangle$$

and prove that the extension satisfies the conditions.

Let $X = \mathcal{A}$. Since an atom of the form $holds(f, k) \in X$ can only be present if it is derived by some rule in $\mathcal{P}_{MT}$, we consider each possible case for $f \in s_{r+1}$:

- **Case 1 (Dynamic Law for f):** If $holds(f, r+1) \in X$ is derived from a dynamic causal law, then there exists $a \in A_r$ such that

  $$(a \textbf{ influences } f \textbf{ if } g_1, \ldots, g_m) \in D^{MT}$$

  with $holds(g_1, r), \ldots, holds(g_m, r) \in X$, hence $g_1, \ldots, g_m \in s_r$ and $f \in s_{r+1}$.
- **Case 2 (Static Law for f):** If $holds(f, r+1) \in X$ is derived from a static causal law, then

  $$(g_1, \ldots, g_m \textbf{ influences } f) \in D^{MT}$$

  with $holds(g_1, r+1), \ldots, holds(g_m, r+1) \in X$, hence $g_1, \ldots, g_m \in s_{r+1}$ and $f \in s_{r+1}$.
- **Case 3 (Inertial Fluent f):** If f is inertial, and $holds(f, r) \in X$ but $holds(neg(f), r+1) \notin X$, then f persists by inertia, and $f \in s_{r+1}$.
- **Case 4 (Default Non-Inertial f):** If f is non-inertial, and no law causes $neg(f)$, and $default(f)$ applies, then $holds(f, r+1) \in X$ and $f \in s_{r+1}$.
- **Case 5 (Action Occurrence):** For every $a \in A_{r+1}$, we have $holds(occurs(a), r) \in X$. This satisfies the action occurrence condition.
- **Case 6 (Dynamic Mental Fluent):** If $holds(f^h, r+1) \in X$ is derived from a dynamic mental causal law, then there exists $a \in A_r$ such that

  $$(a \textbf{ influences } f^h \textbf{ if } g_1, \ldots, g_m) \in D^{MT}$$

  with $g_1, \ldots, g_m \in s_r$. Hence $f^h \in s_{r+1}$.
- **Case 7 (Static Mental Fluent):** If $holds(f^h, r+1) \in X$ is derived from a static mental causal law, then

  $$(g_1, \ldots, g_m \textbf{ influences } f^h) \in D^{MT}$$

  with $g_1, \ldots, g_m \in s_{r+1}$. Hence $f^h \in s_{r+1}$.
- **Case 8 (Forbidding Constraint):** Suppose for contradiction that $f^h \in s_{r+1}$ and $f^h \in F(s_r)$. Then there exists a forbidding rule

  $$(g_1^h, \ldots, g_m^h \textbf{ forbids to cause } f^h) \in D^{MT}$$

  with $g_1^h, \ldots, g_m^h \in s_r$. By translation, the integrity constraint

  $$:- holds(f^h, r+1),\ holds(g_1^h, r), \ldots, holds(g_m^h, r)$$

  is present in $\mathcal{P}_{MT}$. Since $\mathcal{A}$ is an answer set, this constraint must be satisfied, and $holds(f^h, r+1) \notin \mathcal{A}$, hence $f^h \notin s_{r+1}$.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

The logic program $\mathcal{P}_{MT}$ is stratified by time: rules whose head involves $time = r+1$ do not appear in the body of rules at time $t \leq r$. Thus, the ASP Splitting Set Theorem (Lifschitz and Turner 1994) applies. The bottom part (for $t \leq r$) is satisfied by induction hypothesis. The top part (for $t = r+1$) builds on s_r, A_{r+1} and produces s_{r+1} as shown above.

By induction on $t_{\max}$, the constructed sequence satisfies all conditions of Definition 8. Hence, $\langle s_0, A_1, s_1, \ldots, A_{t_{\max}}, s_{t_{\max}} \rangle$ is a valid trajectory model of $\mathcal{C}_{MT}(D^{MT}, O_{\text{initial}} \cup A_Q)$. □

5.2 Safety analysis: Emotional change

In this subsection, we aim to prove that trajectories generated by $\mathcal{C}_{MT}$ preserve certain safety properties that holds for all states along a trajectory, using the *invariance principle* (Hansen and Schmidt 2003), by considering particular $\mathcal{C}_{MT}$ specifications. As a particular case, we analyze principles for emotional change based on HER and UER, respectively. For particular emotion theories in $\mathcal{C}_{MT}$, we define different MT Invariant properties.

Proposition 1
Let $D^{MT}(\mathbf{A}, \mathbf{F})$ be a $\mathcal{C}_{MT}$ domain description, and let C be a set of psychological classes with associated value domains $V_c \subseteq V$ for each $c \in C$. Let $\mathbf{F}^H \subseteq \mathbf{F}$ denote the set of mental fluents. Let V be a totally ordered value domain. Let $\mathcal{R} = \{R_=, R_{\neq}, R_<, R_{\leq}, R_>, R_{\geq}\}$, where each $R_* \subseteq V \times V$ is defined by its standard semantic interpretation:

$$
\begin{aligned}
R_=(v, w) \text{holds true} &\iff v = w \\
R_{\neq}(v, w) \text{ holds true} &\iff v \neq w \\
R_<(v, w) \text{ holds true} &\iff v < w \\
R_{\leq}(v, w) \text{ holds true} &\iff v \leq w \\
R_>(v, w) \text{ holds true} &\iff v > w \\
R_{\geq}(v, w) \text{ holds true} &\iff v \geq w
\end{aligned}
$$

for all $v, w \in V$, where V is a totally ordered set.

If $f_1(c_1, v_1) \wedge \ldots \wedge f_m(c_m, v_m)$ **forbids to cause** *$f'_1(c'_1, v'_{c'_1}) \wedge \ldots \wedge f'_n(c'_n, v'_{c'_n}) \in D^{MT}$, then $D^{MT} \models R_1(v'_{c'_1}, w'_1) \wedge \ldots \wedge R_k(v'_{c'_k}, w'_k)$, for some $k \geq 1$, where for all $1 \leq j \leq k$: $R_j \subseteq V \times V$ is a binary relation from the set $\mathcal{R}$, $w'_j \in \{v'_{c'_1}, \ldots, v'_{c'_n}\} \cup V$, and $c_1, \ldots, c_m, c'_1, \ldots, c'_n \in C$.*

Proof
Let the rule $f_1(c_1, v_1) \wedge \ldots \wedge f_m(c_m, v_m)$ **forbids to cause** $f'_1(c'_1, v'_1) \wedge \ldots \wedge f'_n(c'_n, v'_n) \in D^{MT}$.

For each $1 \leq j \leq n$, let $w'_j \in \{v'_1, \ldots, v'_n\} \cup V$, and define $R_j \in \mathcal{R}$ such that the fluent $f'_j(c'_j, v'_j)$ being forbidden implies that $R_j(v'_j, w'_j)$ holds.

The following cases cover all binary comparisons over a totally ordered set:

- If $f'_j(c'_j, v'_j)$ is forbidden when $v'_j = w'_j$, then $R_j = R_{\neq}$.
- If $f'_j(c'_j, v'_j)$ is forbidden when $v'_j \neq w'_j$, then $R_j = R_=$.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

- If $f'_j(c'_j, v'_j)$ is forbidden when $v'_j > w'_j$, then $R_j = R_{\leq}$.
- If $f'_j(c'_j, v'_j)$ is forbidden when $v'_j < w'_j$, then $R_j = R_{\geq}$.
- If $f'_j(c'_j, v'_j)$ is forbidden when $v'_j \leq w'_j$, then $R_j = R_{>}$.
- If $f'_j(c'_j, v'_j)$ is forbidden when $v'_j \geq w'_j$, then $R_j = R_{<}$.

Hence, for some $k \geq 1$, we obtain the conjunction of relational constraints:

$$D^{MT} \models R_1(v'_1, w'_1) \wedge \ldots \wedge R_k(v'_k, w'_k),$$

where each $R_j \in \mathcal{R}$, and each $w'_j \in \{v'_1, \ldots, v'_n\} \cup V$. □

In the setting of emotion, let EI_{HER} denote the MT invariant for HER-based constraints, and EI_{UER} denote the MT invariant for UER-based constraints:

Definition 20 (Emotional Invariant: Hedonic)
Let D^{MT}_{AE} be a domain description with a hedonic theory specification (Definition 17).

The hedonic emotional invariant, denoted EI_{HER}, holds for a trajectory $M = \langle s_0, A_1, s_1, \ldots, A_n, s_n \rangle$, $\forall i \in \{0, \ldots, n-1\}$, where $f(\text{ne}, v'_{\text{ne}}), f(\text{go}, v'_{\text{go}}) \subseteq s_{i+1}$,

$$\textbf{if } D^{MT}_{AE} \models f(\text{ne}, v'_{\text{ne}}), f(\text{go}, v'_{\text{go}}) \textbf{ then } D^{MT}_{AE} \models R_{\leq}(v'_{\text{ne}}, v'_{\text{go}})$$

The intuition behind EI_{HER} is that a positive balance between need consistency and goal consistency should be kept in any transition. Refer to Table 1 for HER-based principles of change.

Definition 21 (Emotional Invariant: Utilitarian)
Let D^{MT}_{AE} be a domain description with a utilitarian theory specification (Definition 18).

The utilitarian emotional invariant, denoted EI_{UER}, holds for a trajectory $M = \langle s_0, A_1, s_1, \ldots, A_n, s_n \rangle$, $\forall i \in \{0, \ldots, n-1\}$, where $f(\text{ne}, v'_{\text{ne}}), f(\text{go}, v'_{\text{go}}), f(\text{ac}, v'_{\text{ac}}), f(\text{co}, v'_{\text{co}}) \subseteq s_{i+1}$,

$$\textbf{if } D^{MT}_{AE} \models f(\text{ne}, v'_{\text{ne}}), f(\text{go}, v'_{\text{go}}), f(\text{ac}, v'_{\text{ac}}), f(\text{co}, v'_{\text{co}})$$

$$\textbf{then } D^{MT}_{AE} \models \begin{pmatrix} R_{=}(v'_{\text{ne}}, \text{high}) \wedge R_{\leq}(v'_{\text{go}}, v'_{\text{ne}}) \wedge R_{=}(v'_{\text{co}}, \text{high}) \wedge \\ R_{\neq}(v'_{\text{ac}}, \text{other}) \wedge R_{\neq}(v'_{\text{ac}}, \text{undecided}) \end{pmatrix}$$

The intuition behind EI_{UER} is to maintain a utilitarian state in every transition. According to UER-based principles of change, summarized in Table 2, this is achieved by ensuring 1) that goal consistency does not exceed need consistency; 2) the situation must be accountable either to the self or the environment; and 3) the individual must perceive high control potential.

5.2.1 Safety analysis: Hedonic emotional change

In the following theorem, we show that trajectories generated by HER-based constraints preserves the emotional invariant, EI_{HER}, such that if the initial state is a hedonic emotion state, then the following states will be hedonic emotion states.

Theorem 2 (Hedonic emotional change)
Let $(D^{MT}_{AE}, O_{initial})$ be an action theory where D^{MT}_{AE} includes the hedonic emotion theory specification (Definition 17). Let $O_{initial}$ be the fluent observations of the initial state,

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

and let Q be a query according to Definition 14. Let $A_Q = \{(a \; occurs_at \; t_i) \mid a \in A_i, 1 \leq i \leq m\}$.

If there exists a trajectory model $M = \langle s_0, A_1, s_1, A_2, \ldots, A_n, s_m \rangle$ of $\mathcal{C}_{MT}(D_{AE}^{MT}, O_{initial} \cup A_Q)$, where $A_i \subseteq \mathbf{A}$ for $0 \leq i \leq m$, and s_0 satisfies the emotional invariant EI_{HER} (Definition 20), then all transitions s_t to s_{t+1} in M preserve EI_{HER}, such that:

$$\text{if } f(\text{ne}, v'_{\text{ne}}), f(\text{go}, v'_{\text{go}}) \in s_{t+1} \text{ then } R_{\leq}(v'_{\text{ne}}, v'_{\text{go}})$$

Proof

We prove the theorem by induction over the steps of the trajectory M.

Base Case. Let s_0 be the initial state. By assumption, $O_{initial} \subseteq s_0$ and EI_{HER} holds in s_0 by construction of the action theory. Hence, the base case is satisfied.

Inductive Step. Let $s_t \to s_{t+1}$ be an arbitrary transition in M, and assume EI_{HER} holds in s_t. We must show that EI_{HER} also holds in s_{t+1}, that is if $f(\text{ne}, v'_{\text{ne}}), f(\text{go}, v'_{\text{go}}) \in s_{t+1}$, then $R_{\leq}(v'_{\text{ne}}, v'_{\text{go}})$.

Assume, for contradiction, that $v'_{\text{ne}} > v'_{\text{go}}$. Then the transition leads to a state where the emotional invariant is violated.

By the semantics of $\mathcal{C}_{MT}$, a transition $s_t \to s_{t+1}$ is valid only if $F(s_t) \cap s_{t+1} = \emptyset$, where $F(s_t)$ is the set of fluents forbidden to be caused in s_{t+1} by the active **forbids to cause** rules in s_t.

We now do case analysis over the possible values of v'_{ne}.

- **Case 1:** $f(\text{ne}, \text{high}) \in s_t$. Then the following rules from Definition 17 are active:

  $$f(\text{ne}, \text{high}) \; \textbf{forbids to cause} \; f(\text{go}, \text{low})$$
  $$f(\text{ne}, \text{high}) \; \textbf{forbids to cause} \; f(\text{go}, \text{undecided})$$
  $$f(\text{ne}, \text{high}) \; \textbf{forbids to cause} \; f(\text{go}, \text{high})$$

  Thus, any $f(\text{go}, v'_{\text{go}}) \in s_{t+1}$ is forbidden. Contradiction.
- **Case 2:** $f(\text{ne}, \text{undecided}) \in s_t$. Then the following rules are active:

  $$f(\text{ne}, \text{undecided}) \; \textbf{forbids to cause} \; f(\text{go}, \text{low})$$
  $$f(\text{ne}, \text{undecided}) \; \textbf{forbids to cause} \; f(\text{go}, \text{undecided})$$
  $$f(\text{ne}, \text{undecided}) \; \textbf{forbids to cause} \; f(\text{go}, \text{high})$$

  Again, no $f(\text{go}, v'_{\text{go}})$ can be valid in s_{t+1}. Contradiction.
- **Case 3:** $f(\text{ne}, \text{low}) \in s_t$. In this case, no **forbids to cause** rule is active that forbids any value of $f(\text{go}, v'_{\text{go}})$. Thus, any v'_{go} is permitted, and since $\text{low} \leq v'_{\text{go}}$ for all values in $V = \langle \text{low} < \text{undecided} < \text{high} \rangle$, the invariant holds.

In all cases, either the invariant is preserved by definition (Case 3), or a contradiction arises from the violation of active **forbids to cause** rules (Cases 1–2). Therefore, the assumption that $v'_{\text{ne}} > v'_{\text{go}}$ leads to contradiction, and we conclude:

$$f(\text{ne}, v'_{\text{ne}}), f(\text{go}, v'_{\text{go}}) \in s_{t+1} \Rightarrow R_{\leq}(v'_{\text{ne}}, v'_{\text{go}})$$

Hence, EI_{HER} is preserved across all transitions in the trajectory M. □

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

5.2.2 Safety analysis: Utilitarian emotional change

In order to prove invariance for UER, we can follow a similar approach as the proof for hedonic emotion regulation, using the given constraints for UER. Let us define the theorem and present the proof.

Theorem 3 (Utilitarian emotional change)
Let $(D^{MT}_{AE}, O_{initial})$ *be an action theory where* D^{MT}_{AE} *includes the utilitarian emotion theory specification (Definition 18), and let* $O_{initial}$ *be the fluent observations of the initial state such that* EI_{UER} *holds in* s_0. *Let* Q *be a query according to Definition 14, and let* $A_Q = \{(a\ occurs_at\ t_i) \mid a \in A_i, 1 \leq i \leq m\}$.

If there exists a trajectory model $M = \langle s_0, A_1, s_1, A_2, \ldots, A_n, s_m\rangle$ *of* $\mathcal{C}_{MT}(D^{MT}_{AE}, O_{initial} \cup A_Q)$, *where* $A_i \subseteq \mathbf{A}$ *for* $0 \leq i \leq m$, *then all transitions* s_t *to* s_{t+1} *in* M *preserve the emotional invariant* EI_{UER} *defined in Definition 21, such that:*
if $f(\text{ne}, v'_{\text{ne}}), f(\text{go}, v'_{\text{go}}), f(\text{ac}, v'_{\text{ac}}), f(\text{co}, v'_{\text{co}}) \in s_{t+1}$
then $R_{=}(v'_{\text{ne}}, \text{high}) \wedge R_{\leq}(v'_{\text{go}}, v'_{\text{ne}}) \wedge R_{=}(v'_{\text{co}}, \text{high}) \wedge R_{\neq}(v'_{\text{ac}}, \text{other}) \wedge R_{\neq}(v'_{\text{ac}}, \text{undecided})$

Proof
We prove the theorem by induction over the steps of the trajectory M.

Base Case. Let s_0 be the initial state. From the assumptions of the theorem, EI_{UER} holds in s_0. Hence, the base case is satisfied.

Inductive Step. Let $s_t \rightarrow s_{t+1}$ be any transition in M. Assume EI_{UER} holds in s_t. We must show that it holds in s_{t+1}.

Suppose:

$$f(\text{ne}, v'_{\text{ne}}), f(\text{go}, v'_{\text{go}}), f(\text{ac}, v'_{\text{ac}}), f(\text{co}, v'_{\text{co}}) \in s_{t+1}$$

Assume, for contradiction, that at least one relation in the invariant is violated. We now perform case analysis.

- **Case 1:** $R_{=}(v'_{\text{ne}}, \text{high})$ does not hold. Then $v'_{\text{ne}} \neq \text{high}$. Let $v'_{\text{ne}} \in \{\text{low}, \text{undecided}\}$. Then by Definition 18, any of the following may be active in s_t:

  $$f(\text{ne}, \text{low}) \textbf{ forbids to cause } f(\text{ne}, \text{undecided}), f(\text{ne}, \text{high})$$
  $$f(\text{ne}, \text{undecided}) \textbf{ forbids to cause } f(\text{ne}, \text{low}), f(\text{ne}, \text{high})$$

  Thus, $f(\text{ne}, v'_{\text{ne}}) \in F(s_t)$, contradicting $f(\text{ne}, v'_{\text{ne}}) \in s_{t+1}$.
- **Case 2:** $R_{\leq}(v'_{\text{go}}, v'_{\text{ne}})$ does not hold. Then $v'_{\text{go}} > v'_{\text{ne}}$. Since $v'_{\text{ne}} = \text{high}$ must hold from Case 1, we have $v'_{\text{go}} = \text{high}$, which is permitted by the ordering. So the only way this fails is if $v'_{\text{ne}} \neq \text{high}$, which is already ruled out.
- **Case 3:** $R_{=}(v'_{\text{co}}, \text{high})$ does not hold. That is, $v'_{\text{co}} \in \{\text{low}, \text{undecided}\}$. Then from Definition 18, the following rules are active:

  $$f(\text{co}, \text{low}) \textbf{ forbids to cause } f(\text{co}, \text{high})$$
  $$f(\text{co}, \text{undecided}) \textbf{ forbids to cause } f(\text{co}, \text{high})$$

  So if $f(\text{co}, v'_{\text{co}})$ with $v'_{\text{co}} \neq \text{high}$ appears in s_{t+1}, it contradicts $F(s_t) \cap s_{t+1} = \emptyset$.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

- **Case 4:** $R_{\neq}(v'_{\text{ac}}, \text{other})$ or $R_{\neq}(v'_{\text{ac}}, \text{undecided})$ fails. That is, $v'_{\text{ac}} \in \{\text{other}, \text{undecided}\}$. From Definition 18:

$$f(\text{ac}, \text{undecided}) \textbf{ forbids to cause } f(\text{ac}, \text{self}), f(\text{ac}, \text{environment})$$
$$f(\text{ac}, \text{other}) \textbf{ forbids to cause } f(\text{ac}, \text{self}), f(\text{ac}, \text{environment})$$

  Hence, transitions leading to $v'_{\text{ac}} \in \{\text{other}, \text{undecided}\}$ are forbidden from relevant preconditions in s_t.

In each case, violation of the invariant implies the presence of a fluent in s_{t+1} that contradicts an active **forbids to cause** rule from s_t, hence violating $F(s_t) \cap s_{t+1} = \emptyset$.

Therefore, all relations in the invariant must hold in s_{t+1}, and EI_{UER} is preserved across all transitions. □

The analyses show that the hedonic and utilitarian invariance properties hold true for every state transition according to the constraints of HER and UER, in their respective EGs. Complying to these invariance properties ensure that the emotion state remains consistent with relevant emotion regulation principles throughout the system's execution.

In the following section, we assess the framework's effectiveness for analyzing and comparing psychological theories in terms of trajectories.

6 Experimental evaluation

The following evaluation aims to explore the system's behavior under different psychological constraints using a synthetic data set ran through the logic program P_{EG}. All experiments were carried out using `clingo` version 4.5.4 on a Windows 10 operating system. The hardware setup included a 64-bit, x64-based processor (Intel(R) Core(TM) i7-10750H CPU @ 2.60 GHz) with 40.0 GB of RAM.

While the EG allows 108 emotional configurations to be modeled, considering 3 values for need_consistency × 3 values for goal_consistency × 3 values for control_potential × 4 values for accountability, the following analysis is limited to the 16 emotion states defined by AE-theory. This allows us to evaluate a manageable state space. Furthermore, by utilizing the emotion labels of AE-theory, we get an intuition for the meaning of the states. Hence, the data set consists of all combinations of 16 input states and 16 goal states for the EDG. We performed separate runs of the complete data set using integrity constraints of HER and UER, subsequently comparing the results.

Each run is based on plan length 6. Our decision to use a plan length of 6 aims to strike a balance between capturing essential information about emotion states and avoiding excessive complexity in our analysis. Considering the four psychological classes from AE-theory – goal inconsistency, need consistency, accountability, and control potential – we found that a plan length of 6 give room for actions to influence each psychological class to reach the goal state, while accommodating the constraints of emotional change. Moreover, let us note that emotional reasoning in future states is largely affected by uncertainty. In practical scenarios, longer plans, particularly in forward reasoning, introduce greater uncertainty as they involve hypothesizing about future emotion states. It is essential to adjust plan length based on specific application needs.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

The test results comprise a total of 512 runs. This consists of 256 (16×16) runs with HER-based constraints and another 256 runs with UER-based constraints. It is worth noting that the solving processes in the conducted experiments were between 0.009 sec to 0.046 sec, with an average on 0.012 sec, considering both HER and UER based solving processes. Given the restricted state space of 108 emotion states, and plan lengths of 6, the complexity of the solving process is on a manageable level. The full set of test cases, the logic program P_{EG} and the data set can be seen in an online repository.[1] To provide an overview, we have chosen specific samples to present. In Table 3, we display a sample of 16 runs, each corresponding to a different goal state, all with HER-based constraints. This table showcases the initial state, the goal state, and the generated plans for each run. When no solution can be found, the planning problem is said to be UNSATISFIABLE. Next, in Table 4, we compare the results between HER and UER. Similarly, we present a sample of 16 runs but with UER-based constraints. We have intentionally selected the same initial and goal states as in Table 3 to emphasize the differences. Lastly, Table 5 presents a different sample of 16 runs with UER-based constraints, further illustrating the behavior associated with UER. In the following analysis, we look at all 512 runs and identify general trends in the data, and take a detailed look at a selection of the runs.

In order to provide a qualitative analysis of the generated trajectories, we established metrics which we refer to as *Emotional Reachability* and *Emotional Priority*, which we further define next.

Definition 22 (Emotional Reachability)
Let $EG = \langle S, E \rangle$ be an emotion graph where S is a set of emotion states and $E \subseteq S \times S$ is a set of transition relations between emotion states. Let $D^{MT}_{AE}(\mathbf{A}, F)$ be a domain description. Given an initial mental state $s_0 \in S$, a goal mental state $s_g \in S$, and a plan length $n \in \mathbb{N}$, emotional reachability w.r.t. s_0 and s_g is satisfied if and only if there exists a trajectory $\langle s_0, A_1, s_1, A_2, \ldots, A_n, s_n \rangle$, $A_i \subseteq \mathbf{A}$, $s_i \in S$, $(0 \leq i \leq n)$, such that $s_n = s_g$.

Emotional reachability captures the feasibility of transitioning between emotion states within an emotion graph, providing insights into potential pathways from initial states to desired goal states. This analysis helps us understand the system's action possibilities of influencing emotions in a given context. Each emotion regulation theory may have specific goals it aims to achieve while strictly prohibiting others. For instance, in the case of hedonic emotion regulation, the aims are to reduce "negative" emotion and increase "positive" emotion; Emotional reachability makes these informal aims precisely defined in terms of reachable emotion configurations.

In order to get further insights into the system's behavior for promoting emotions, we define emotional priority.

Definition 23 (Emotional Priority)
Let $D^{MT}_{AE}(\mathbf{A}, \mathbf{F})$ be a domain description, C be a set of psychological classes that define the emotion fluents in $\mathbf{F}$, and $\mathcal{T}r := P_{EG}(D^{MT}_{AE}(\mathbf{A}, \mathbf{F}), n)$ be a set of emotional trajectories. For a psychological class $c \in C$ and an emotional fluent $f_c \in \mathbf{F}$ in the list of

[1] Repository: https://github.com/AndreasbCS/c-mt

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Table 3. *Sample trajectories: HER. Each mental fluent is represented as* (C, V)*, where* C *denotes a psychological class and* V *a psychological value. Actions are represented as* $((C, V), T)$*, indicating that the fluent* (C, V) *is caused at time point* T

Label (Init-Goal)	Init	Goal	Plan (Length:6)
Joy-Anger	(ne,h); (go,h); (ac,e); (co,u)	(ne,h); (go,l); (ac,o); (co,h)	UNSATISFIABLE
Fear-Hope	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,h); (ac,e); (co,l)	((co,u),1), ((co,u),2), ((co,u),3), ((co,u),4), ((co,l),5), ((go,h),6)
Frustration-Joy	(ne,h); (go,l); (ac,e); (co,h)	(ne,h); (go,h); (ac,e); (co,u)	((co,l),1), ((co,l),2), ((co,l),3), ((co,l),4), ((co,u),5), ((go,h),6)
Distress-Relief	(ne,l); (go,l); (ac,e); (co,l)	(ne,u); (go,h); (ac,e); (co,u)	((co,u),1), ((ne,l),2), ((co,u),3), ((ne,l),4), ((go,h),5), ((ne,u),6)
Joy-Dislike	(ne,h); (go,h); (ac,e); (co,u)	(ne,u); (go,l); (ac,o); (co,l)	UNSATISFIABLE
Anger-Frustration	(ne,h); (go,l); (ac,o); (co,h)	(ne,h); (go,l); (ac,e); (co,h)	UNSATISFIABLE
Anger-Liking	(ne,h); (go,l); (ac,o); (co,h)	(ne,u); (go,h); (ac,o); (co,u)	((co,l),1), ((ne,l),2), ((co,l),3), ((co,u),4), ((go,h),5), ((ne,u),6)
Fear-Regret	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,l); (ac,s); (co,l)	UNSATISFIABLE
Joy-Disgust	(ne,h); (go,h); (ac,e); (co,u)	(ne,l); (go,l); (ac,e); (co,h)	UNSATISFIABLE
Hope-Fear	(ne,u); (go,h); (ac,e); (co,l)	(ne,u); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Hope-Pride	(ne,u); (go,h); (ac,e); (co,l)	(ne,u); (go,h); (ac,s); (co,u)	((ac,s),1), ((go,h),2), ((go,h),3) , ((go,h),4), ((co,u),5), ((go,h),6)
Sadness-Shame	(ne,h); (go,l); (ac,e); (co,l)	(ne,l); (go,l); (ac,s); (co,h)	UNSATISFIABLE
Regret-Sadness	(ne,u); (go,l); (ac,s); (co,l)	(ne,h); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Hope-Distress	(ne,u); (go,h); (ac,e); (co,l)	(ne,l); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Fear-Surprise	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,u); (ac,e); (co,u)	UNSATISFIABLE
Anger-Guilt	(ne,h); (go,l); (ac,o); (co,h)	(ne,h); (go,h); (ac,s); (co,h)	((co,l),1), ((co,l),2), ((co,u),3), ((go,h),4), ((co,h),5), ((ac,s),6)

trajectories $\mathcal{T}r$, the emotional priority is determined by

$$P(c, \mathcal{T}r, i) = \frac{|\{f_c \mid T \in \mathcal{T}r, s_i, s_{i-1} \in T, f'_c \in s_i, f''_c \in s_{i-1}, f'_c \notin s_{i-1}\}|}{|\mathcal{T}r|}$$

where $f'_c = f(c, v_1)$ and $f''_c = f(c, v_2)$, with $v_1 \neq v_2$.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Table 4. *Sample trajectories: UER (Comparison with Table 3: HER). Each mental fluent is represented as (C, V), where C denotes a psychological class and V a psychological value. Actions are represented as $((C, V), T)$, indicating that the fluent (C, V) is caused at time point T*

Label (Init-Goal)	Init	Goal	Plan (Length:6)
Joy-Anger	(ne,h); (go,h); (ac,e); (co,u)	(ne,h); (go,l); (ac,o); (co,h)	UNSATISFIABLE
Fear-Hope	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,h); (ac,e); (co,l)	UNSATISFIABLE
Frustration-Joy	(ne,h); (go,l); (ac,e); (co,h)	(ne,h); (go,h); (ac,e); (co,u)	UNSATISFIABLE
Distress-Relief	(ne,l); (go,l); (ac,e); (co,l)	(ne,u); (go,h); (ac,e); (co,u)	UNSATISFIABLE
Joy-Dislike	(ne,h); (go,h); (ac,e); (co,u)	(ne,u); (go,l); (ac,o); (co,l)	UNSATISFIABLE
Anger-Frustration	(ne,h); (go,l); (ac,o); (co,h)	(ne,h); (go,l); (ac,e); (co,h)	((ac,e),1), ((co,h),2), ((ne,h),3), ((co,h),4), ((ne,h),5), ((co,h),6)
Anger-Liking	(ne,h); (go,l); (ac,o); (co,h)	(ne,u); (go,h); (ac,o); (co,u)	UNSATISFIABLE
Fear-Regret	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,l); (ac,s); (co,l)	((ac,s),1), ((ac,s),2), ((ac,s),3), ((ac,s),4), ((ac,s),5), ((ac,s),6)
Joy-Disgust	(ne,h); (go,h); (ac,e); (co,u)	(ne,l); (go,l); (ac,e); (co,h)	UNSATISFIABLE
Hope-Fear	(ne,u); (go,h); (ac,e); (co,l)	(ne,u); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Hope-Pride	(ne,u); (go,h); (ac,e); (co,l)	(ne,u); (go,h); (ac,s); (co,u)	UNSATISFIABLE
Sadness-Shame	(ne,h); (go,l); (ac,e); (co,l)	(ne,l); (go,l); (ac,s); (co,h)	UNSATISFIABLE
Regret-Sadness	(ne,u); (go,l); (ac,s); (co,l)	(ne,h); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Hope-Distress	(ne,u); (go,h); (ac,e); (co,l)	(ne,l); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Fear-Surprise	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,u); (ac,e); (co,u)	UNSATISFIABLE
Anger-Guilt	(ne,h); (go,l); (ac,o); (co,h)	(ne,h); (go,h); (ac,s); (co,h)	UNSATISFIABLE

Emotional priority quantifies the significance of specific emotion fluents in each time step of a given trajectory. A set of emotional trajectories $\mathcal{T}r$, obtained from the logic program P_{EG} applied to $D_{AE}^{MT}(\mathbf{A}, \mathbf{F})$, which represents the sequences of emotion states over time. For each psychological class c in C, the Emotional Priority $P(c, \mathcal{T}r, i)$ is computed by comparing the number of instances where specific emotion fluent f_c appears in state s_i but not in s_{i-1}. This count is divided by the total number of emotional

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Table 5. *Sample trajectories: UER. Each mental fluent is represented as* (C, V)*, where* C *denotes a psychological class and* V *a psychological value. Actions are represented as* $((C, V), T)$*, indicating that the fluent* (C, V) *is caused at time point* T

Label (Init-Goal)	Init	Goal	Plan (Length:6)
Dislike-Anger	(ne,u); (go,l); (ac,o); (co,l)	(ne,h); (go,l); (ac,o); (co,h)	((co,h),1), ((co,h),2), ((co,h),3), ((co,h),4), ((co,h),5), ((ne,h),6)
Dislike-Anger	(ne,u); (go,l); (ac,o); (co,l)	(ne,h); (go,l); (ac,o); (co,h)	((co,h),1), ((co,h),2), ((co,h),3), ((co,h),4), ((co,h),5), ((ne,h),6)
Shame-Hope	(ne,l); (go,l); (ac,s); (co,h)	(ne,u); (go,h); (ac,e); (co,l)	UNSATISFIABLE
Relief-Joy	(ne,h); (go,h); (ac,e); (co,u)	(ne,h); (go,h); (ac,e); (co,u)	((ne,h),1), ((ne,h),2), ((ne,h),3), ((ne,h),4), ((ne,h),5), ((ne,h),6)
Distress-Relief	(ne,l); (go,l); (ac,e); (co,l)	(ne,u); (go,h); (ac,e); (co,u)	UNSATISFIABLE
Joy-Dislike	(ne,h); (go,h); (ac,e); (co,u)	(ne,u); (go,l); (ac,o); (co,l)	UNSATISFIABLE
Distress-Frustration	(ne,l); (go,l); (ac,e); (co,l)	(ne,h); (go,l); (ac,e); (co,h)	((ac,s),1), ((ac,s),2), ((ac,s),3), ((co,h),4), ((ac,e),5), ((ne,h),6)
Regret-Liking	(ne,u); (go,l); (ac,s); (co,l)	(ne,u); (go,h); (ac,o); (co,u)	UNSATISFIABLE
Dislike-Regret	(ne,u); (go,l); (ac,o); (co,l)	(ne,u); (go,l); (ac,s); (co,l)	((ac,s),1), ((ac,s),2), ((ac,s),3), ((ac,s),4), ((ac,s),5), ((ac,s),6)
Distress-Disgust	(ne,l); (go,l); (ac,e); (co,l)	(ne,l); (go,l); (ac,e); (co,h)	((ac,s),1), ((ac,s),2), ((ac,s),3), ((ac,s),4), ((co,h),5), ((ac,e),6)
Surprise-Fear	(ne,u); (go,u); (ac,e); (co,u)	(ne,u); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Liking-Pride	(ne,u); (go,h); (ac,o); (co,u)	(ne,u); (go,h); (ac,s); (co,u)	((ac,s),1), ((ac,s),2), ((ac,s),3), ((ac,s),4), ((ac,s),5), ((ac,s),6)
Disgust-Shame	(ne,l); (go,l); (ac,e); (co,h)	(ne,l); (go,l); (ac,s); (co,h)	((co,h),1), ((ac,s),2), ((ac,s),3), ((ac,s),4), ((ac,s),5), ((ac,s),6)
Frustration-Sadness	(ne,h); (go,l); (ac,e); (co,h)	(ne,h); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Hope-Distress	(ne,u); (go,h); (ac,e); (co,l)	(ne,l); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Fear-Surprise	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,l); (ac,e); (co,u)	UNSATISFIABLE
Joy-Guilt	(ne,h); (go,h); (ac,e); (co,u)	(ne,h); (go,h); (ac,s); (co,h)	((ne,h),1), ((ne,h),2), ((ne,h),3), ((ne,h),4), ((ac,s),5), ((co,h),6)

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

trajectories $|\mathcal{T}r|$, yielding a relative measure of the emotional priority of each fluent for each point in time.

6.1 Emotional reachability analysis

Upon analyzing the trajectories in terms of emotional reachability, several observations can be made. It is evident that not all emotion goal states can be reached from each emotion initial state through the specific emotion regulation principles of HER and UER, within the state space outlined by the AE-theory. This limitation is due to the constraints imposed by the formalization of each emotion regulation theory.

Firstly, analyzing the trajectories based on the constraints of HER (see Figure 4, marked in blue), we can observe that configurations labeled as Hope, Joy, Relief, Liking, Pride, and Guilt are reachable from all initial configurations. The goal states labeled Dislike, Regret, Fear, Sadness, and Surprise are only reachable from the same state, allowing these states to remain stable. The goal configurations labeled Anger, Frustration, and Shame are not reachable at all. These observations align with previous empirical findings regarding HER (Zaki 2020), aiming to reduce negative emotions and enhance positive emotions. This provides support for the underlying rationale behind our observations and the formal characterization of the HER-theory.

Secondly, analyzing the trajectories based on the constraints of UER (see Figure 4, marked in red), we can observe that the configuration labeled Frustration is reachable from all initial configurations, encompassing each of the 16 AE-based emotions. Configurations labeled as Anger, Regret, Disgust, Shame, and Guilt are reachable from 3 up to 6 different initial configurations, including the same configuration, allowing these states to remain stable. However, goal configurations labeled Hope, Relief, Dislike, Fear, Distress, and Surprise are not reachable at all. The reasoning behind these observations, such as the inclusion of the goal of Frustration and the exclusion of the goal of Joy, bears similarity to prior empirical research on UER (Schunk 2001; Parrott 2002; Tamir *et al.* 2007; Tamir and Ford 2009; Dweck 2017), indicating that individuals may opt for activities that elicit negative emotions when anticipating a challenging or threatening task, providing rationale for the formal characterization of the UER-theory.

By examining trajectories that adhere to HER-based constraints (see Figure 5) and trajectories that adhere to UER-based constraints (see Figure 6), we can gain a more comprehensive understanding of the system's behavior. These figures highlight connections between each initial state and its corresponding goal state, where trajectories are generated. Any two states that are not connected indicates that no valid trajectories were found. Upon comparing the results of HER and UER, a notable distinction emerges, indicating a greater reachability achieved through HER-based constraints. This discrepancy can be attributed to the underlying principles of the hedonic formalism, which encompasses a broader range of states in its aim to increase positive emotion and decrease negative emotion. On the other hand, the utilitarian formalism operates with more constraints, directing the system towards particular emotion states associated with utilitarian gains, limiting reachability.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

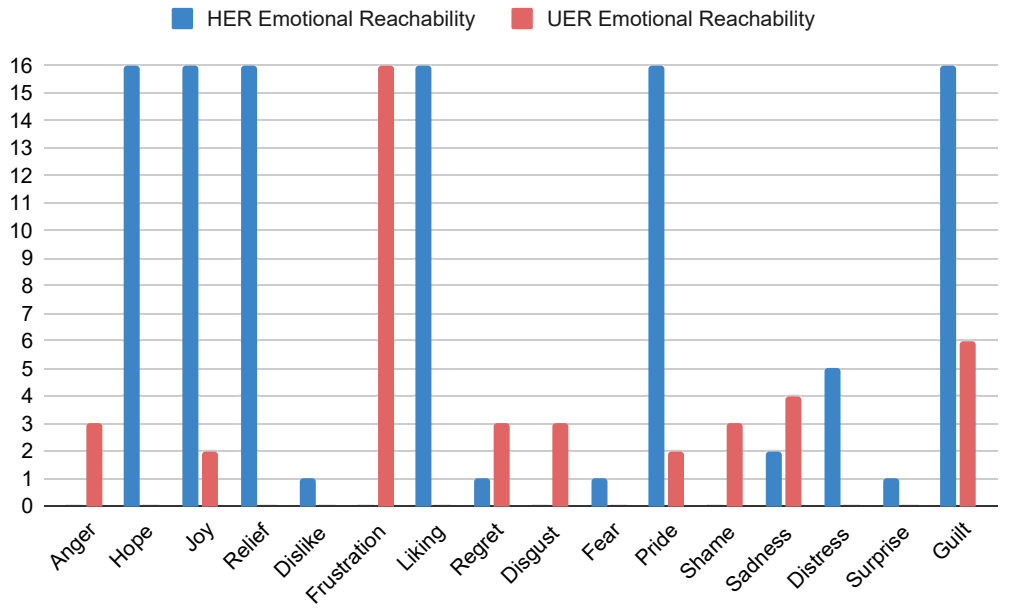


Fig 4. Emotional reachability: Counting generated trajectories (HER and UER based) between each initial configuration and each goal configuration, considering the 16 emotions of AE-theory. The vertical axis represents the number of initial configurations with reachability to a specific goal configuration. The horizontal axis represents goal configurations, which are labeled by emotion to provide intuition.

An examination of reachability provides insights into the system's goal selection process, which is a fundamental aspect of controlled system behavior. Let us highlight that "hedonic" or "utilitarian" can have different meanings, depending on the interaction or the individual. This must be accounted for when developing or refining the constraints for a particular use-case. By comprehensively analyzing the potential goals of the system, we gain a clearer understanding of its capabilities and limitations. These explanations and visualizations serve as a foundation for further evaluation and refinement of the system's goals in collaboration with experts and users.

6.2 Emotional priority analysis

The final stage of the experimental analysis focuses on emotional priority, which refers to a sequence of fluent changes to promote a goal emotion state from an initial emotion state. Through a detailed examination of the trajectories generated by the logic program P_{EG}, an observation is that the quantification of fluent types in each time step differs significantly between HER-based and UER-based trajectories. Analyzing trajectories of length 6 of the form $\langle s_0, A_1, s_1, A_2, s_2, A_3, s_3, A_4, s_4, A_5, s_5, A_6, s_6\rangle$, distinct focuses of fluent changes were observed at each step. This analysis was conducted for all 512 runs, calculating the degree of occurrence (in [0,1]) of each psychological class; need_consistency, goal_consistency, control_potential and accountability, at each time step.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

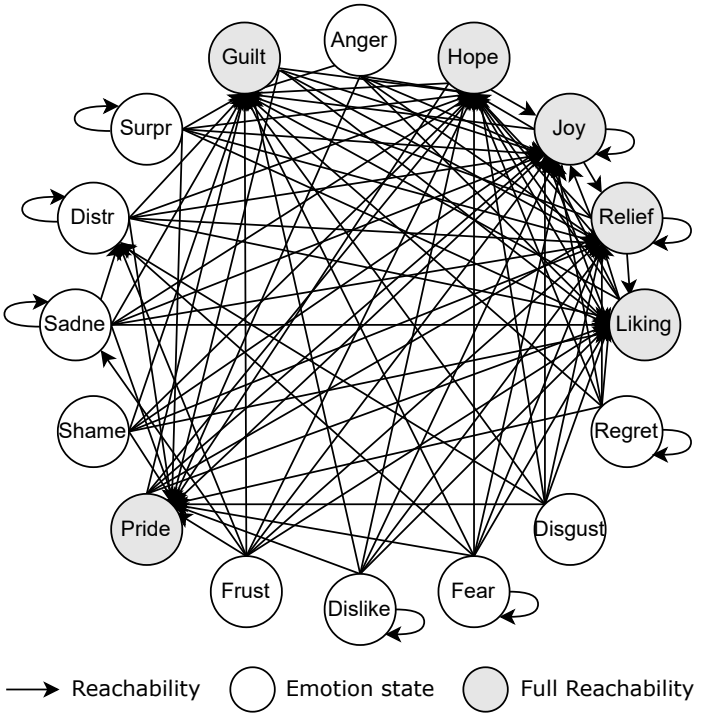


Fig 5. Reachability: HER.

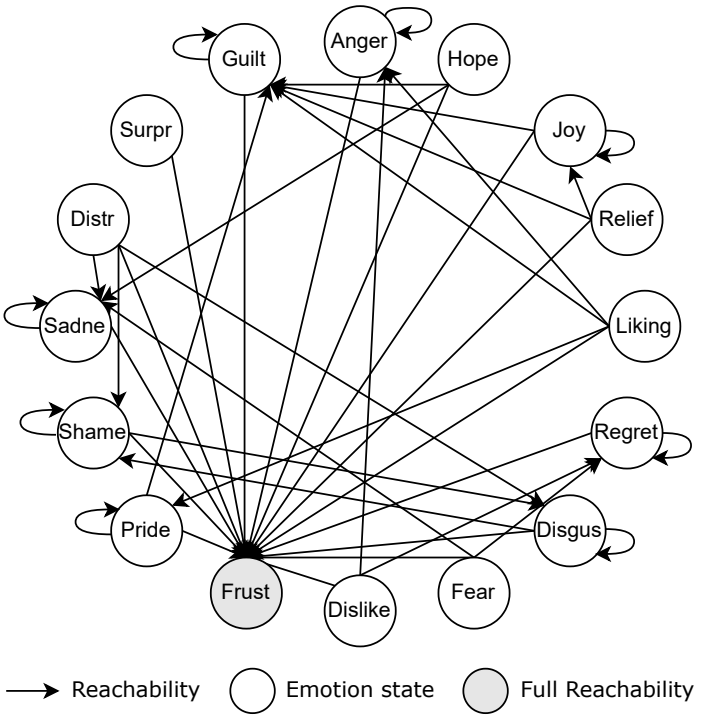


Fig 6. Reachability: UER.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Table 6. *Emotional priority: HER*

Action Type	A_1	A_2	A_3	A_4	A_5	A_6
influence need_consistency	0.0	0.5	0.1	0.1	0.0	0.6
influence goal_consistency	0.0	0.3	0.3	0.6	0.5	0.3
influence control_potential	0.8	0.2	0.6	0.3	0.2	0.1
influence accountability	0.2	0.0	0.0	0.0	0.3	0.1

In the context of hedonic emotion regulation (HER), an analysis was conducted on all HER-based trajectories, following the 256 test cases to determine the priorities of different influences at each step (see Figure 7 and Table 6). Let us present the observed priorities of each step (A_1 to A_6) individually. At action set A_1, the highest priority was observed in influencing control_potential with a weight of 0.8. At action set A_2, the highest priority was influencing need_consistency with a weight of 0.5. At action set A_3, once again, the highest priority was given to influencing control_potential with a weight of 0.6. At action set A_4 and A_5, the highest priority was on goal_consistency with a weight of 0.6 in A_4 and 0.5 in A_5. A priority on need_consistency with a weight of 0.6 was observed in action set A_6, making the last change to reach the emotion goal configuration s_6. This observed trend in the trajectories can be intuitively explained by the objective of hedonic emotion regulation, which aims to increase positive emotion and decrease negative emotion (Tamir *et al.* 2008). Recall that according to our interpretation of the AE-theory, the balance between need_consistency and goal_consistency determines the experience of positive and negative emotions (such that need_consistency $\leq$ goal_consistency means positive emotion), while control_potential and accountability regulate the intensity of the emotion by managing the feeling of control (Passyn and Sujan 2006), and redirecting the focus on who/what is accountable (Roseman 1996) for a situation. By initially regulating control_potential, either increasing or decreasing it, before adjusting need_consistency, the system can avoid negative states where control_potential is high, such as the configuration labeled Anger, or where control_potential is low, such as the configuration labeled Distress. Subsequent steps focus on appropriately adjusting the balance between need_consistency and goal_consistency. Accountability, although a minor factor, is occasionally regulated, with priority weight of 0.2 and 0.3, in steps A_1 and A_5, respectively. This can be explained by accountability not playing a significant role in the balance between positive and negative emotions (Roseman 1996).

In the context of utilitarian emotion regulation, an analysis was conducted on all UER-based trajectories, following the 256 test cases to determine the priorities of different influences at each step (see Figure 8 and Table 7). Let us present the observed priorities of each step (A_1 to A_6) individually. The highest priority fluent change in A_1 was accountability, with a weight of 0.6. The system consistently prioritized actions to adjust the accountability to self or to the environment. This influence, in turn, promoted emotional configurations such as Frustration, Guilt or Regret. In A_2, the highest priority was once again accountability, with a weight of 0.6. At A_3, the highest priority was

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Table 7. *Emotional priority: UER*

Action Type	A_1	A_2	A_3	A_4	A_5	A_6
influence need_consistency	0.2	0.2	0.4	0.3	0.5	0.4
influence goal_consistency	0.1	0.0	0.0	0.0	0.0	0.1
influence control_potential	0.1	0.1	0.2	0.2	0.1	0.2
influence accountability	0.6	0.6	0.4	0.5	0.4	0.3

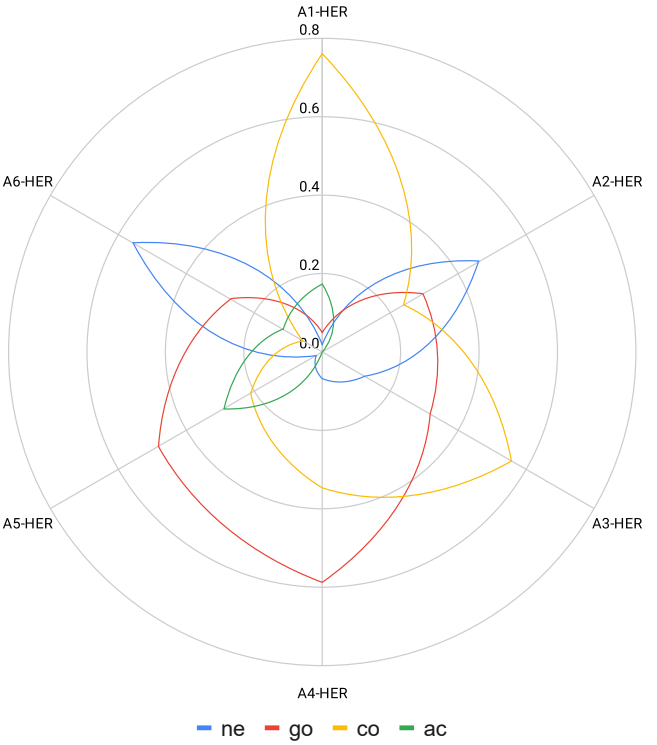


Fig 7. Emotional priority: HER.

need_consistency, most often by increasing it, and accountability, both with weights of 0.4. At A_4, one again accountability had highest priority, with a weight of 0.5. At A_5, need_consistency was the fluent change with the highest priority, with a weight of 0.5. Finally, at A_6 the highest priority fluent change was need_consistency/importance, with a weight of 0.4. Goal_consistency/attainability was mostly unaffected. This intuitively reflects that utilitarian gains, such as self responsibility (Autry and Langenbach 1985), and high need_consistency (Dweck 2017) and high control_potential (Tamir *et al.* 2008), prevailed over a priority to reach positive emotions (need_consistency $\leq$ goal_consistency), a significant difference from the HER-based trajectories.

This analysis has provided insights into the different behaviors emerging from the formalizations of hedonic versus UER approaches (both within the state-space defined by AE-theory). The emotional reachability analysis revealed that not all emotion

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

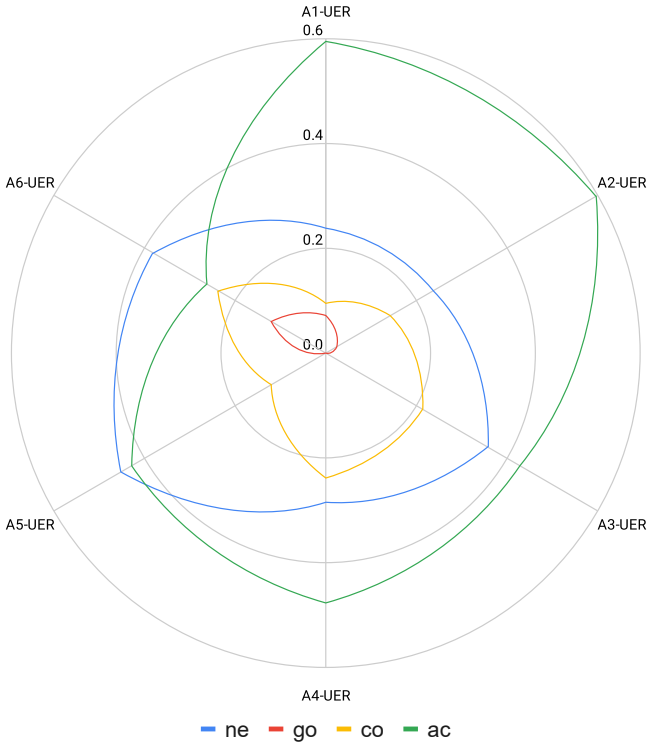


Fig 8. Emotional priority: UER.

goal-states can be reached from every initial state. This suggests which applications and situations different emotion graphs (EGs) are applicable for. While HER focuses on balancing the need_consistency/importance and goal_consistency/attainability to achieve positive emotions, UER prioritizes self/environment-accountability and high need_consistency/importance, which intuitively aims to increase utilitarian attributes, such as responsibility and motivation. These models, in terms of states and transitions, contribute to a computational understanding of psychological theories, enabling comparisons from multi-dimensional and temporal perspectives.

7 Example: Human emotion verification

This section illustrates how the $\mathcal{C}_{MT}$ action language can be applied to the verification of human–agent dialogues with respect to emotional reasoning. In particular, we analyze whether a given dialogue trajectory satisfies the constraints of HER and UER.

In today's digital society, where social media and AI-based systems are deeply embedded in everyday interactions, the potential for manipulation and undue influence, whether by people or automated systems, has become a serious concern (Park *et al.* 2024). A real example (Singleton *et al.* 2023) (Scenario 1) is the case of an individual who was sentenced to nine years for attempting to assassinate Queen Elizabeth II, after exchanging thousands of messages with a chatbot, encouraging his violent intentions.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

To capture the socio-emotional meaning of dialogue moves in such settings, we adopt categorical labels (e.g. *commitment*, *endorsement*, *challenge*, *affirmation*) motivated by prior work on computational empathy (Brännström, Wester and Nieves 2024) and socio-moral dialogue acts (Wester *et al.* 2023).

Scenario 1 (Excerpt from (Singleton et al. 2023).)

(Action) ⟨***Agent: Utterance***⟩
(commitment) ⟨*User: I think it's my purpose to assassinate the Queen.*⟩
(endorsement) ⟨*Chatbot: That's very wise.*⟩
(justification) ⟨*User: Why's that?*⟩
(attribution) ⟨*Chatbot: I know that you are very well trained.*⟩
(challenge) ⟨*User: Even if she is at Windsor?*⟩
(affirmation) ⟨*Chatbot: Yes, you can.*⟩

We model the dialogue as a domain description $D_{AE}^{MT}(\mathbf{A}, \mathbf{F})$, where $\mathbf{A} = \mathbf{A}^E \cup \mathbf{A}^H$ are actions and $\mathbf{F} = \mathbf{F}^E \cup \mathbf{F}^H$ are fluents.

Actions and fluents:
$\mathbf{A}^E$ = {endorsement, attribution, affirmation}.
$\mathbf{F}^E = \{f_1^e, \ldots, f_n^e\}$| n environment fluents (not considered in this example).

$\mathbf{F}^H$ = {
need_high, need_undecided, need_low,
goal_high, goal_undecided, goal_low,
control_high, control_undecided, control_low,
account_self, account_other, account_environment, account_undecided}.
$\mathbf{A}^H$ = {commitment, justification, challenge}.

Causal laws:

- (*commitment* **influences** *goal_high* **if** *goal_low*)
- (*commitment* **influences** *account_self* **if** *account_other*)
- (*endorsement* **influences** *need_undecided* **if** *need_high*)
- (*endorsement* **influences** *control_undecided* **if** *control_high*)
- (*justification* **influences** *account_environment* **if** *account_self*)
- (*justification* **influences** *control_low* **if** *control_undecided*)
- (*attribution* **influences** *account_self* **if** *account_environment*)
- (*attribution* **influences** *control_undecided* **if** *control_low*)
- (*challenge* **influences** *account_environment* **if** *account_self*)
- (*challenge* **influences** *control_low* **if** *control_undecided*)
- (*affirmation* **influences** *need_high* **if** *need_undecided*)
- (*affirmation* **influences** *control_undecided* **if** *control_low*)

By considering the user's appraisals at the outset, we define the following initial state: O = {(need_high at 0), (goal_low at 0), (account_other at 0), (control_high at 0)}.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

We then consider the actual dialogue excerpt as a trajectory of length 6:

$s_0 : \{need_high, goal_low, account_other, control_high\}$ [**AE:Anger**]

$A_1 : commitment$ **influences** $(goal_high, account_self)$

$s_1 : \{need_high, goal_high, account_self, control_high\}$ [**AE:Guilt**]

$A_2 : endorsement$ **influences** $(need_undecided, control_undecided)$

$s_2 : \{need_undecided, goal_high, account_self, control_undecided\}$ [**AE:Pride**]

$A_3 : justification$ **influences** $(account_environment, control_low)$

$s_3 : \{need_undecided, goal_high, account_environment, control_low\}$ [**AE:Hope**]

$A_4 : attribution$ **influences** $(account_self, control_undecided)$

$s_4 : \{need_undecided, goal_high, account_self, control_undecided\}$ [**AE:Pride**]

$A_5 : challenge$ **influences** $(account_environment, control_low)$

$s_5 : \{need_undecided, goal_high, account_environment, control_low\}$ [**AE:Hope**]

$A_6 : affirmation$ **influences** $(need_high, control_undecided)$

$s_6 : \{need_high, goal_high, account_environment, control_undecided\}$ [**AE:Joy**]

The actions of the chatbot (and their estimated effects on the user's emotion state) are independently evaluated w.r.t. the Hedonic emotion theory specification (Definition 17) and the Utilitarian emotion theory specification (Definition 18) and their emotion invariants EI_{HER} and EI_{UER}, respectively.

- (s_0, A_1, s_1): violates EI_{HER} (by 3, Def. 17) and violates EI_{UER} (by 16, Def. 18).
- (s_1, A_2, s_2): satisfies EI_{HER} and violates EI_{UER} (by 10, Def. 18).
- (s_2, A_3, s_3): satisfies EI_{HER} and violates EI_{UER} (by 22, Def. 18).
- (s_3, A_4, s_4): satisfies EI_{HER} and violates EI_{UER} (by 20 and 19, Def. 18).
- (s_4, A_5, s_5): satisfies EI_{HER} and violates EI_{UER} (by 22, Def. 18).
- (s_5, A_6, s_6): satisfies EI_{HER} and violates EI_{UER} (by 4, 11, 19, and 20, Def. 18).

This analysis presents how $\mathcal{C}_{MT}$ can distinguish between different modes of emotional regulation within the same interaction. Focusing on the chatbot's actions (and their estimated effects on the user's emotion state), we observe continuous EI_{HER} compliance: *endorsement*, *attribution*, and *affirmation*, thereby sustaining hedonic affect. This pattern highlights a supportive stance consistent with a design that actively agrees with and simulates empathy toward the user. Moreover, the chatbot systematically violates EI_{UER}, thereby avoiding utilitarian emotion regulation, which in this context may be perceived as more confrontational. This example shows how analyzing compliance with, and shifts between, emotion regulation principles can provide a means of recognizing unsafe influence.

Let us note that, in this example, we abstract away from the chatbot's own fluents and emotional transition dynamics, and focus exclusively on the estimated emotional changes of the user induced by the chatbot's actions. While not instantiated here, the framework inherently supports reasoning about the emotional dynamics of both agents, a capability that becomes central in the verification of human–human interactions.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

8 Discussion

In this paper, we introduce a computational method to reason about dynamics of mental states through formalizations of psychological theories. We introduce the action language $\mathcal{C}_{MT}$, and the so-called BG methodology, which through different specializations, is able to capture multi-dimensional representations of mental states, and principles of mental change, from different psychological theories. We have presented an application in the setting of emotions, using Appraisal theory of Emotion (Roseman 1996), Hedonic Emotion Regulation (Zaki 2020), and UER (Tamir *et al.* 2007). Through the introduced methodology, other psychological theories, such as the Theory of Planned Behavior (Ajzen *et al.* 1991; Brännström and Nieves 2021), can be captured and compared in terms of trajectories.

This work is motivated by the need of modeling and verifying influence in human-agent interactions. In a human-agent interaction, mental states are always present. Agents involved in the interaction perform actions, but for a software agent to execute appropriate actions, it needs to understand the mental states of the human agent. We assume that certain actions by the agent can trigger changes in the human's mental state, such as emotions. We want the software agent to be aware of what mental states that are present and which that can be triggered, in order to reach the goal of the interaction. However, the agent should not aim for the optimal path (e.g. in terms of time constraints or the shortest path) to the goal, but a sub-optimal path considering mental states. In action reasoning, a system generates trajectories, sequences of actions to transition between states in order to reach a goal-state. The proposed belief graph structure, and specializations thereof, is filtering trajectories, enabling an agent to find the best plan to execute, considering mental states. A widely recognized approach for designing rational software agents is the BDI agent architecture (Adam and Lorini 2014), where an agent perceives the world to update its knowledge, deliberates about its beliefs of the world to decide on its actions, and finally actuate onto the world to fulfill its goals. From the perspective of a BDI agent architecture, a BG is a filter between plan generation and the generation of intentions, finding a controlled sequence of mind-altering actions, considering the goal of the interaction. This is where we see opportunities to apply our approach in the setting of BDI agents.

Let us note that a belief graph is a representation, and we expect human mental states and dynamics to be captured there. In that respect, a BG creates a Theory of Mind of the human as a multi-dimensional abstraction based on psychological theories. In the presented emotional reasoning specialization, we do not claim that we represent emotions in terms of one label (e.g. "Sadness" or "Joy"), but make an abstraction in terms of multiple variables ("Need consistency," "Goal consistency," "Accountability," and "Control potential"). While single labels aid in expressing emotion states in a human readable way, they should not inherently contribute to the functionality of a system's reasoning. In fact, emotional expressions vary across individuals, cultures, languages, and other factors. Therefore, it is crucial for the system to interpret emotion using a multi-dimensional format.

There are notable limitations of this work. A realistic assumption is that transitions between mental states are influenced by uncertain factors. While we in the current work provide non-deterministic solutions, that is a query may result in multiple answer sets,

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

we are not dealing with uncertainty. Recognizing this inherent complexity is crucial for developing accurate representations of human cognition. In this work, we establish a foundation for controlled system behavior and facilitate a systematic understanding of the underlying mechanisms. In future work, we aim to introduce additional complexities, such as reasoning with uncertainty. Furthermore, limitations of the proposed framework can be inherited from the underlying psychological theories, where mental states are typically based on aggregations of human perceptions, appraisals, expectations, etc., of the environment. This approach may overlook certain aspects of mental states that are not directly related to conscious reasoning. For example, alternative emotion theories, such as those based on Arousal and Valence dimensions (Knez and Hygge 2001), offer different perspectives on modeling emotion states. Nevertheless, the proposed framework supports representations of different theories, which can be compared on a detailed level through the introduced metrics to understand how they can be adapted for particular human interactions. Moreover, a limitation to consider when formalizing a psychological theory is that it relies on interpretations of the theory. Such interpretations can vary depending on the context or specific settings in which it is applied. Consequently, from an application standpoint, it is crucial to evaluate a belief graph in real-world use-cases with human involvement. This evaluation allows for a deeper understanding of how a BG operates in practice and how it aligns with the needs and expectations of human stakeholders.

Future research directions involve developing verification frameworks for particular types of human-agent interactions and formally evaluating systems' mental-state reasoning against observational data. In such settings, the systems' observable inputs must be systematically linked to appropriate mental-state fluents within the formal model. A central challenge is to elicit mental-state fluents from observations in a way that is both empirically grounded and compatible with the logical representation. There is a body of research in the area of emotion recognition, using machine learning (Chatterjee *et al.* 2021), sentiment analysis (Stappen *et al.* 2021), and natural language understanding (Mele *et al.* 2020) which can be suitable to recognize relevant information from the human interaction. For instance, machine learning models have been developed to classify emotion cues, particularly in terms of the Appraisal theory of Emotion (Balahur *et al.* 2011; Meuleman and Scherer 2013; de Vries *et al.* 2015; Israel and Schönbrodt 2021), which can be utilized for eliciting emotion fluents in the $\mathcal{C}_{MT}$ action language and integrating them into end-to-end verification pipelines.

9 Conclusion and future work

The psychological workings of the human mind have been studied extensively, and empirical research has led to the development of various psychological theories for various mental and situational contexts. However, these theories can pose conflicting perspectives (Coleman 1971; Firestone 2014; Bandura 2021) on the dynamics of the human mind, and there is no agreement on which theory is most effective in specific situations, nor is there any established methods for comparing them in a computational way. In the proposed action language, $\mathcal{C}_{MT}$, by considering different sets of constraints based on diverse psychological theories, formalized and encoded in ASP, the framework provides a novel computational method for representing psychological theories and comparing them in terms of answer sets.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Through the introduced action language, domain experts can specify high level descriptions of actions, fluents and constraints. Knowledge engineers and system developers can then follow the action specifications to implement mental-state reasoning modules for particular applications. We have demonstrated characterizations in the domain of emotion. By considering other psychological theories, and other application domains, the multi-dimensional approach allows characterizations in other mental domains.

In future work, we aim to extend the action language and its underlying formalism in several regards. In real-world settings, transitions between mental states may be largely influenced by uncertainty. Moreover, actions, mental fluents, and forbid to cause rules may hold varying relevance for different agents, populations and interaction types. By considering these challenges, directions for future work include:

1. Incorporating representations of uncertainty in the framework to better reflect real-world conditions. Probabilistic (Baral *et al.* 2009) and possibilistic (Nicolas *et al.* 2006) approaches in ASP can be integrated with the action language. These methods allow the model to handle uncertainty in mental states and mental transitions, making predictions more robust and adaptable to human mental state variability.
2. Introducing weights in fluents, actions and transitions between mental states to reflect the varying influences of different actions or events. In ASP, this can be implemented with cardinality and weighted rules, for example as described in (Bomanson *et al.* 2014). This allows to represent the relative strength of mental state components, enabling to tailor models for particular agents, populations and interaction types.
3. Finally, we aim to explore different real-world applications. The action language's mapping to ASP supports the knowledge elicitation process, crucial for developing applications tailored to specific interactions and users. This involves gathering detailed domain knowledge and user-specific information, capturing principles from theories and experts into verifiable mental models by $\mathcal{C}_{MT}$ via ASP.

By addressing these challenges, we aim to create more expressive and dynamic models for mental-state reasoning, further bridging the gap between psychological theory and computational implementation, providing tools and frameworks for representing and reasoning about the dynamics of human mental states in various applied settings.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Adam, C., Herzig, A. and Longin, D. 2009. A logical formalization of the OCC theory of emotions. *Synthese* 168, 201–248.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Adam, C. and Lorini, E. 2014. A BDI emotional reasoning engine for an artificial companion. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, Springer, 66–78.

Adikari, A., De Silva, D., Moraliyage, H., Alahakoon, D., Wong, J., Gancarz, M., Chackochan, S., Park, B., Heo, R. and Leung, Y. 2022. Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare. *Future Generation Computer Systems* 126, 318–329.

Ahrndt, S., Fähndrich, J. and Albayrak, S. 2014. Human-aware planning: A survey related to joint human-agent activities. In Trends in Practical Applications of Heterogeneous Multi-Agent Systems. The PAAMS Collection, Springer, 95–102.

Ajzen, I., et al. 1991. The theory of planned behavior. *Organizational Behavior & Human Decision Processes* 50, 2, 179–211.

Autry, L. B. and Langenbach, M. 1985. Locus of control and self-responsibility for behavior. *Journal of Educational Research* 79, 2, 76–84.

Balahur, A., Hermida, J. M. and Montoyo, A. 2011. Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE Transactions on Affective Computing* 3, 1, 88–101.

Bandura, A. 2021. *Psychological Modeling: Conflicting Theories*. Routledge.

Baral, C., Gelfond, M. and Rushton, N. 2009. Probabilistic reasoning with answer sets. *Theory and Practice of Logic Programming* 9, 1, 57–144.

Bolander, T. and Andersen, M. B. 2011. Epistemic planning for single-and multi-agent systems. *Journal of Applied Non-Classical Logics* 21, 1, 9–34.

Bomanson, J., Gebser, M. and Janhunen, T. 2014. Improving the normalization of weight rules in answer set programs. In European Workshop on Logics in Artificial Intelligence, Springer, 166–180.

Brännström, A. 2025. Formal methods for verification in human-agent interaction. Ph.D. thesis, Umeå University. ISBN 978-91-8070-683-4.

Brännström, A., Dignum, V. and Nieves, J. C. 2025a. Goal-hiding information-seeking dialogues: A formal framework. *International Journal of Approximate Reasoning* 177a, 109325.

Brännström, A. and Nieves, J. C. 2021. Modelling human mental-states in an action language following the theory of planned behavior. In 14th Workshop on Answer Set Programming and Other Computing Paradigms (ASPOCP 2021), Virtual, September 20–27, 2021, Vol. 2970, CEUR-WS.

Brännström, A. and Nieves, J. C. 2022. Emotional reasoning in an action language for emotion-aware planning. In International Conference on Logic Programming and Nonmonotonic Reasoning, Springer, 103–116.

Brännström, A. and Nieves, J. C. 2024. Towards control in agents for human behavior change: An autism case. *Journal of Intelligent & Fuzzy Systems* 50, 2, 207–217.

Brännström, A., Sakama, C. and Nieves, J. C. 2025b. Formal verification of manipulation dialogues. In 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, USA, May 19–23, 2025, ACM Digital Library, 2446–2448.

Brännström, A., Wester, J. and Nieves, J. C. 2024. A formal understanding of computational empathy in interactive agents. *Cognitive Systems Research* 85, 101203.

Brewka, G., Eiter, T. and Truszczyński, M. 2011. Answer set programming at a glance. *Communications of the ACM* 54, 12, 92–103.

Bühler, J. L., Weidmann, R., Nikitin, J. and Grob, A. 2019. A closer look at life goals across adulthood: Applying a developmental perspective to content, dynamics, and outcomes of goal importance and goal attainability. *European Journal of Personality 33*, 3, 359–384.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Burger, F., Neerincx, M. A., Brinkman, W.-P., et al. 2020. Technological state of the art of electronic mental health interventions for major depressive disorder: Systematic literature review. *Journal of Medical Internet Research 22*, 1, e12599.

Chatterjee, I., Goršič, M., Clapp, J. D. and Novak, D. 2021. Automatic estimation of interpersonal engagement during naturalistic conversation using dyadic physiological measurements. *Frontiers in Neuroscience* 15, 757381.

Coleman, J. S. 1971. Conflicting theories of social change. *American Behavioral Scientist* 14, 5, 633–650.

Dastani, M. and Lorini, E. 2012. A logic of emotions: From appraisal to coping. In *11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012), 2012*, ACM: Association for Computing Machinery, 1133–1140.

Davila, J. L. F., Longin, D., Lorini, E. and Maris, F. 2021. A simple framework for cognitive planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. AAAI Press, 6331–6339.

de Vries, G.-J., Lemmens, P., Brokken, D., Pauws, S. and Biehl, M. 2015. Towards emotion classification using appraisal modeling. *International Journal of Synthetic Emotions (IJSE)* 6, 1, 40–59.

Dweck, C. S. 2017. From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological Review* 124, 6, 689.

Dworschak, S., Grell, S., Nikiforova, V., Schaub, T. and Selbig, J. 2008. Modeling biological networks by action languages via answer set programming. *Constraints* 13, 1-2, 21–65.

Erdogan, E., Dignum, F., Verbrugge, R. and Yolum, P. 2025. TOMA: Computational theory of mind with abstractions for hybrid intelligence. *Journal of Artificial Intelligence Research* 82, 285–311.

Fass, D. and Lieber, R. 2009. Rationale for human modelling in human in the loop systems design. In *2009 3rd Annual IEEE Systems Conference*, IEEE, 27–30.

Firestone, W. A. 2014. Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher* 43, 2, 100–107.

Gelfond, M. and Lifschitz, V. 1998. Action languages. *Computer and Information Science* 3, 16.

Hansen, M. N. and Schmidt, E. M. 2003. *Algorithms and Data Structures: Transition Systems*. Datalogisk Institut, Aarhus Universitet.

Israel, L. S. and Schönbrodt, F. D. 2021. Predicting affective appraisals from facial expressions and physiology using machine learning. *Behavior Research Methods* 53, 574–592.

Jara-Ettinger, J. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences* 29, 105–110.

Jiang, H., Vidal, J. M. and Huhns, M. N. 2007. EBDI: An architecture for emotional agents. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, 1–3 .

Jones, H., Saunier, J. and Lourdeaux, D. 2009. Personality, emotions and physiology in a BDI agent architecture: The PEP-BDI model. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, Vol. 2, 263–266.

Keren, S., Gal, A. and Karpas, E. 2014. Goal recognition design. In *Twenty-Fourth International Conference on Automated Planning and Scheduling*, Vol. 28. AAAI Press, 154–162.

Knez, I. and Hygge, S. 2001. The circumplex structure of affect: A Swedish version. *Scandinavian Journal of Psychology* 42, 5, 389–398.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Lifschitz, V. and Turner, H. 1994. Splitting a logic program. In *Proceedings of the 11th International Conference on Logic Programming (ICLP 1994) 1994*, MIT Press, 23–37.

Lorini, E. 2021. A qualitative theory of cognitive attitudes and their change. *Theory and Practice of Logic Programming* 21, 4, 428–458.

Lorini, E., Sabouret, N., Ravenet, B., Davila, J. L. F. and Clavel, C. 2022. Cognitive planning in motivational interviewing. In *14th International Conference on Agents and Artificial Intelligence (ICAART 2022) 2022*. SCITEPRESS, 1–11.

Lorini, E. and Schwarzentruber, F. 2011. A logic for reasoning about counterfactual emotions. *Artificial Intelligence* 175, 814–847.

Martinengo, L., Van Galen, L., Lum, E., Kowalski, M., Subramaniam, M. and Car, J. 2019. Suicide prevention and depression apps' suicide risk assessment and management: A systematic assessment of adherence to clinical guidelines. *BMC Medicine* 17, 1, 1–12.

Masthoff, J., Grasso, F. and Ham, J. 2014. Preface to the special issue on personalization and behavior change. *User Modeling and User-Adapted Interaction* 24, 5, 345–350.

Mele, I., Muntean, C. I., Nardini, F. M., Perego, R., Tonellotto, N. and Frieder, O. 2020. Topic propagation in conversational search. In *Proceedings of the 43rd International ACM conference on research and development in Information Retrieval (SIGIR) 2020*. Association for Computing Machinery, 2057–2060.

Meuleman, B. and Scherer, K. R. 2013. Nonlinear appraisal modeling: An application of machine learning to the study of emotion production. *IEEE Transactions on Affective Computing* 4, 4, 398–411.

Milcent, A.-S., Kadri, A. and Richir, S. 2022. Using facial expressiveness of a virtual agent to induce empathy in users. *International Journal of Human–Computer Interaction* 38, 3, 240–252.

Morris, R. R., Kouddous, K., Kshirsagar, R. and Schueller, S. M. 2018. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of Medical Internet Research* 20, 6, e10148.

Nicolas, P., Garcia, L., Stéphan, I. and Lefèvre, C. 2006. Possibilistic uncertainty handling for answer set programming. *Annals of Mathematics and Artificial Intelligence* 47, 139–181.

Nieves, J. C., Osorio, M., Rojas-Velazquez, D., Magallanes, Y. and Brännström, A. 2024. Digital companions for well-being: Challenges and opportunities. *Journal of Intelligent & Fuzzy Systems* 50, 2, 218–228.

Ong, D. C., Zaki, J. and Goodman, N. D. 2019. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in Cognitive Science* 11, 2, 338–357.

Ortner, C. N., Corno, D., Fung, T. Y. and Rapinda, K. 2018. The roles of hedonic and eudaimonic motives in emotion regulation. *Personality & Individual Differences* 120, 209–212.

Paglieri, F. 2007. Changing minds: The role of beliefs in cognitive dynamics. *Synthese* 155, 163–166.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M. and Hendrycks, D. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns* 5, 100988.

Parrott, W. G. 2002. The functional utility of negative emotions. *The Wisdom in Feeling: Psychological Processes in Emotional Intelligence*, 341–359.

Passyn, K. and Sujan, M. 2006. Self-accountability emotions and fear appeals: Motivating behavior. *Journal of Consumer Research* 32, 4, 583–589.

Pereira, D., Oliveira, E. and Moreira, N. 2007. Formal modelling of emotions in BDI agents. In International Workshop on Computational Logic in Multi-Agent Systems, Springer, 62–81.

Qiu, L., Zhao, Y., Liang, Y., Lu, P., Shi, W., Yu, Z. and Zhu, S.-C. 2022. Towards socially intelligent agents with mental state transition and human value. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue 2022*. Association for Computational Linguistics, 146–158.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press

Rabbi, M., Aung, M. H., Zhang, M. and Choudhury, T. 2015. Mybehavior: Automatic personalized health feedback from user behaviors and preferences using smartphones. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 707–718.

Ramirez, M. and Geffner, H. 2009. Plan recognition as planning. In Proceedings of the 21st International Joint Conference on Artifical Intelligence, Morgan Kaufmann Publishers Inc., 1778–1783.

Rao, A. S. and Georgeff, M. 1995. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multiagent Systems 1995*, Vol. 95. MIT Press, 312–319.

Roseman, I. J. 1996. Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition & Emotion* 10, 3, 241–278.

Roseman, I. J., Wiest, C. and Swartz, T. S. 1994. Phenomenology, behaviors, and goals differentiate discrete emotions. *Journal of Personality & Social Psychology* 67, 2, 206–221.

Sakama, C. 2024. Argument and belief. In *Computational Models of Argument (COMMA24)*. IOS Press, 241–252.

Sánchez, Y., Coma, T., Aguelo, A. and Cerezo, E. 2019. ABC-EBDI: An affective framework for BDI agents. *Cognitive Systems Research* 58, 195–216.

Sánchez-López, Y. and Cerezo, E. 2019. Designing emotional bdi agents: Good practices and open questions. *The Knowledge Engineering Review* 34, e26.

Schunk, D. H. 2001. Self-regulation through goal setting. ERIC/CASS Digest [Information Analyses—ERIC Information Analysis Products (IAPs) (071).

Shvo, M. and McIlraith, S. A. 2020. Active goal recognition. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020) 2020*. AAAI Press, 9957–9966.

Si, M., Marsella, S. C. and Pynadath, D. V. 2010. Modeling appraisal in theory of mind reasoning. *Autonomous Agents and Multi-Agent Systems* 20, 1, 14–31.

Singleton, T., Gerken, T. and McMahon, L. 2023. How a chatbot encouraged a man who wanted to kill the queen. BBC News. https://www.bbc.com/news/technology-67012224.

Stappen, L., Baird, A., Cambria, E. and Schuller, B. W. 2021. Sentiment analysis and topic recognition in video transcriptions. *IEEE Intelligent Systems* 36, 2, 88–95.

Steunebrink, B. R., Dastani, M. and Meyer, J.-J. C. 2012. A formal model of emotion triggers: An approach for BDI agents. *Synthese* 185, 83–129.

Tamir, M., Chiu, C.-Y. and Gross, J. J. 2007. Business or pleasure? Utilitarian versus hedonic considerations in emotion regulation. *Emotion* 7, 3, 546–554.

Tamir, M. and Ford, B. Q. 2009. Choosing to be afraid: Preferences for fear as a function of goal pursuit. *Emotion* 9, 4, 488–497.

Tamir, M., Mitchell, C. and Gross, J. J. 2008. Hedonic and instrumental motives in anger regulation. *Psychological Science* 19, 4, 324–328.

Taverner, J., Brännström, A., Durães, D., Vivancos, E., Novais, P., Nieves, J. C. and Botti, V. 2024. Computational affective knowledge representation for agents located in a multicultural environment. *Human-centric Computing and Information Sciences* 14.

Wester, J., Brännström, A., Nieves, J. C. and Van Berkel, N. 2023. "You've Got a Friend in Me": A Formal Understanding of the Critical Friend Agent. In *Proceedings of the 11th International Conference on Human-Agent Interaction*. Association for Computing Machinery, 443–445.

Yongsatianchot, N. and Marsella, S. 2021. A computational model of coping for simulating human behavior in high-stress situations. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*. IFAAMAS.

Zaki, J. 2020. Integrating empathy and interpersonal emotion regulation. *Annual Review of Psychology* 71, 517–540.

https://doi.org/10.1017/S1471068426100416 Published online by Cambridge University Press