Conv-BDI: An Extension of the BDI Framework for Conversational Agents

Nicholas Thomas Walker, Nicolas Wagner, Laetitia Hilgendorf, Stefan Ultes
Natural Language Generation and Dialogue Systems Group
Otto-Friedrich University of Bamberg, Germany

Abstract

With the large and increasing variety of architectures for conversational agent design, there is a need to investigate the necessary elements of practical conversational agents in light of new technologies. To address this need, we introduce a new abstract framework of conversational agents design which we call *Conv-BDI*. The BDI model is a long-established theory of decision-making in artificial agents, which the Conv-BDI model extends to describe the design of conversational agents from traditional symbolic logic-based models or statistical models to more recent LLM-based agents. Specifically, we extend the core BDI model with notions of *Purpose* and *Behavioral Guidelines*, while also elaborating on the role of system actions within this framework. The Conv-BDI model thus provides a framework of intentionality in conversational agents that can be applied to to design of contemporary conversational agents.

1 Introduction

In the years following the deep learning revolution in Natural Language Processing (NLP), the design of conversational agents has seen substantial evolution, e.g. (Shum et al., 2018; Caldarini et al., 2022). Moving beyond earlier rule-based (McTear, 2021) and statistical models (Griol et al., 2008; Ultes et al., 2017), large language models (LLMs) have become a core component of contemporary conversational agents. LLMs have become foundational to many systems with the use of strategies such as in-context learning and prompt engineering (Bommasani et al., 2021), however they do not in and of themselves represent the full breadth of conversational agents (Yi et al., 2024). Navigating the challenges of decision making and conversational *intentionality* in collaboration with humans remains a core question in dialogue systems research (Lin et al., 2024).

In this context, the goal of this paper is to outline a general framework of conversational agent design, which we call **Conv-BDI**. With this framework based upon the well-established Beliefs-Desires-Intention model of autonomous agents (Rao and Georgeff, 1997), this paper investigates the following core questions:

- *What conceptual components are necessary for the design of conversational agents as intelligent, rational agents?*
- *How do these components depend on each other and interact?*

With respect to a conversational agent as an *intelligent agent*, a long-standing topic of research in conversational agents is modelling the dialogue *policy*, or how the agent should decide upon its next action. Actions in dialogue may include linguistic, gestural, auxiliary actions such as API calls, or a mixture of these. Numerous possible actions in a nondeterministic environment must work towards a long term purpose or tasks for which the system is intended (Russell and Norvig, 2016). As noted by Lin et al. (2024): "[Automated systems] may also be able to efficiently reason under uncertainty about the expected value of decision-relevant information, helping them determine what information may be important to share with or request from the user." To capture this ability, we seek to identify conceptual categories of a conversational agent as an autonomous agent directed towards goals, that is, with **intentionality**.

To design rational agents in the sense defined by e.g. (Russell and Norvig, 2016), we seek a framework encompassing the breadth of architectures for conversational agents to encompass an ever-increasing diversity such systems. For this purpose, we begin with the BDI model as a starting point. The BDI model is a long-standing framework used to outline the design of intelligent

Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue, September 3–5, 2025, Bielefeld, Germany.

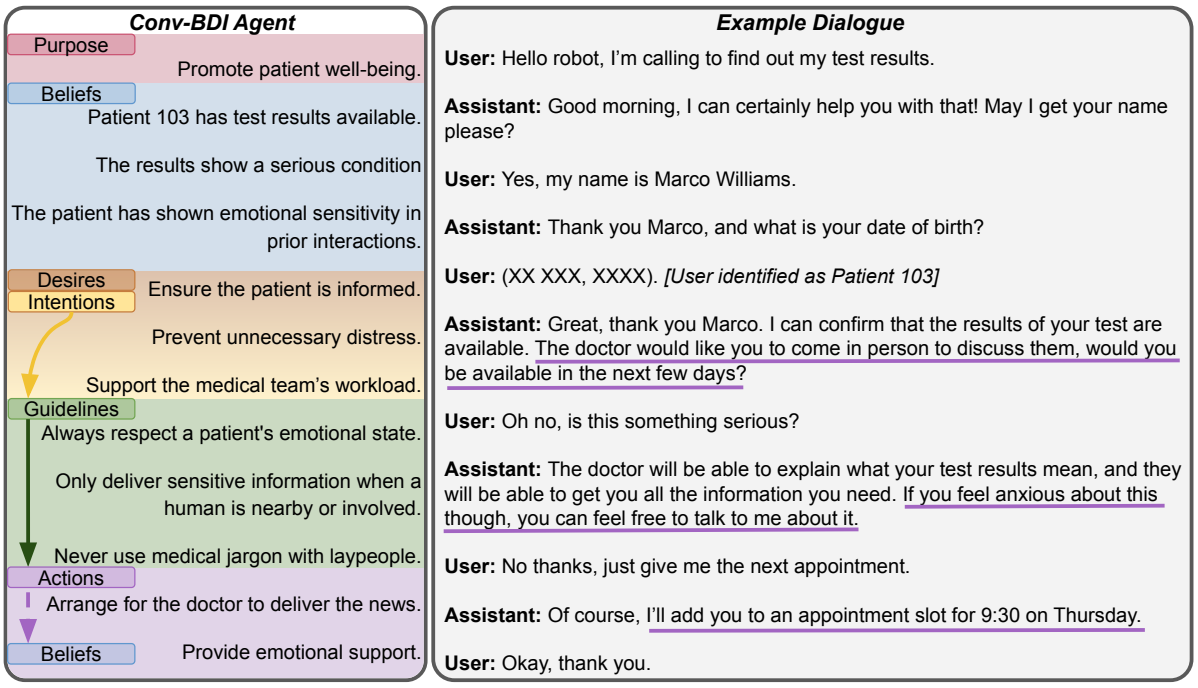

Figure 1: Example of a conversational agent as a healthcare assistant with Conv-BDI components highlighted on the left. This example system has Desires in the context of the Purpose of promoting patient well-being and its Beliefs about the world. Each Desire may be selected as an *Intention* to be carried out with respect to the Guidelines and instantiated by the Actions constrained by the Guidelines, indicated by arrows. The Actions may also update the Beliefs where the Action causes a change in the dialogue state, indicated by a partial arrow.

agents (conversational and otherwise). Building upon BDI, we propose an extension of this model for the design of conversational agents which also incorporates notions of *Purpose*, *Guidelines*, and an extended description of system *Actions* to complement core BDI components. Specifically, these conceptual categories describe the following:

- **Purpose:** The purpose of the agent describes the high level reason for the agent's existence. Conceptually, this is the source of the desires and intentions of the model.
- **Guidelines:** Behavioral Guidelines are the constraints under which the system's intentions and actions should be carried out.
- **Actions:** Actions are fundamental capabilities of the system to effect a result oriented towards an Intention. Actions are performed by the system to create results in the world state (affecting future Beliefs) to fulfill the Intention chosen from Desires derived from the Purpose, within the constraints of Guidelines.

The relation of these components is illustrated in Figure 1, showing the conceptual architecture instantiated as an example conversational agent for healthcare. As a further contribution to the introduction of Conv-BDI, we elaborate on how contemporary LLM-based architectures for conversational agents can be felicitously described within the Conv-BDI framework, as well as earlier agents prior to LLMs. Comparing several architecturally distinct conversational agents, we provide a practical mapping from the theoretical concepts of autonomous agents to current advances in LLM-based conversational agents and show how Conv-BDI characterizes practical implementations of these systems.

2 Related Work

Designing AI systems as rational agents is a continuing topic of research interest (Vetrò et al., 2019). In many respects, the goal of designing a conversational agent is to mimic human behavior. For instance, Cassell et al. (2000) described how characteristics of human-human interaction can serve as the basis for an architecture for designing embodied conversational agents. Meanwhile, BDI has been used as a simulation for human-like decision making in simulations (Adam and Gaudou, 2016). Much previous work on BDI architectures was conducted prior to breakthroughs in

deep learning (Broersen et al., 2005; Holvoet and Valckenaers, 2006). Nonetheless, the BDI model sees continued use for some conversational agents (Ichida and Meneguzzi, 2023).

Extensions to the BDI model have also been considered from perspectives such as emotions and psychology (Sánchez et al., 2019). Other work has begun to investigate ways to incorporate Theory of Mind for agents based on neural architectures (Bortoletto et al., 2024). Recent models such as Deepseek R1 (Guo et al., 2025) have demonstrated impressive capabilities with the integration of expressed "thoughts" leading to the model's output. Even so, LLMs specifically have been observed to lack illocutionary intent in the sense that it is understood in humans: Actions or communications undertaken with the expectation of effecting a change in the world (Rosen and Dale, 2024).

A similar strand of research concerns agentic systems (Shavit et al., 2023). "Agenticness" with regards to autonomous systems relates to the agent's ability to perform goals and tasks with limited direct supervision. Definitions given for agenticness focus on the degree of autonomy of the system and goal complexity, whereas for our theory we focus on the agent's planning capabilities.

3 Conv-BDI: Core Components

In this section, we describe the core elements of Conv-BDI drawn from the established BDI model and their instantiation in contemporary conversational agents. The classic formulation of BDI calls for three components termed *Beliefs*, *Desires*, and *Intentions* (Rao and Georgeff, 1997).

3.1 Beliefs

The first component of Conv-BDI drawn from the classical BDI model is *Beliefs*. The Beliefs of the model are the collection of world knowledge needed for the model to complete its task. Beliefs in a conversational agent are the system's knowledge of the world, and the basis for the system to make decisions and take actions.

A conversational agent must handle a variety of what Russell and Norvig (2016) term "percepts", which are the individual stimuli it has the capability to perceive. Minimally, the system must have an understanding of the immediate dialogue utterance history, otherwise its responses will be incoherent. Further, a system may require access to background information necessary for tasks in a task-oriented system or otherwise general world knowledge in an open-domain system in order to make the responses factual and engaging.

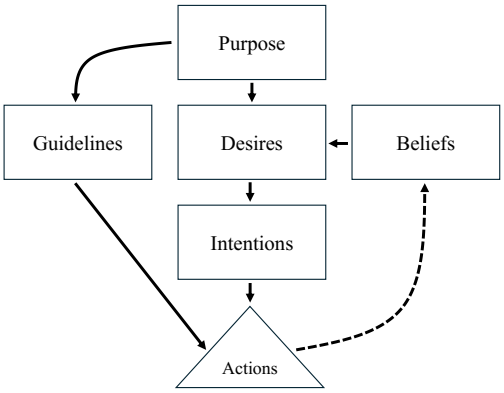

Figure 2: Conceptual dependencies of the Conv-BDI framework. The dashed arrow represents updates to Beliefs from actions, which are optional as all action results need not necessarily be saved in the Beliefs.

In addition, meta-information about the dialogue may optionally be included, including logical forms of dialogue acts or natural language understanding (NLU). With embodied or situated dialogue agents, multi-modal information such as visually perceived objects or the agent's physical position may also be represented. All such elements constitute Beliefs of the system, representing the system's understanding of its context, separate from what it intends to do with this information. The system's Beliefs are an abstract collection of information from potentially heterogeneous sources, depending upon the context, purpose, and practical constraints of the system.

In practical terms, the Beliefs of the system are an *explicitly* represented component a conversational agent. In an end-to-end system, the parameters of the model encode the system's Beliefs. Other systems represent Beliefs in a structured form, e.g. as a knowledge base or dialogue state representations, as in dialogue state tracking tasks (Williams et al., 2016). The Beliefs of a conversational agent in modern LLM-based conversational agents are often made available to the system using Retrieval Augmented Generation (Lewis et al., 2020). Knowledge-grounded conversational agents e.g. Chawla et al. (2024) rely on a structured representation of knowledge that, while external to the LLM specifically, is integral to the function of the system as a conversational agent. In this sense, both the parametric memory

of the LLM and the non-parametric memory retrieved elsewhere jointly constitute the Beliefs of the model. However, at any given moment only certain elements of the wider Beliefs will be relevant for the system to make decisions. Which elements are relevant must be identified in relation to the system's *Desires*.

3.2 Desires

A Desire is any goal that the system might attempt to achieve. Each individual Desire represents a world in which a given set of conditions are fulfilled, e.g. a table at a restaurant has been booked or a window has been closed. Individual Desires may be mutually exclusive with one another. For instance, an embodied agent cannot occupy two places at once. Which Desires are possible to pursue depends on the Purpose and other Beliefs at the current time, as shown in Figure 2. While the Desires of the system represent some aspect of a future world state, they are a subset of that state. That is, there are elements of the system Beliefs outside of the Desire, and the Desire may be a completely novel addition to the Beliefs.

A task-oriented conversational agent specifically aims to accomplish specific goals for the user. For such a system, the Desires are the successful completion of goals provided by the user. A practical example is the user goals in the BPL framework of Zhao et al. (2024), represented as text descriptions. In general, it remains a continuing subject of research to adapt conversational agents to a wider range of domains, characterizable as open-domain conversational agents (Algherairy and Ahmed, 2024). Viewed within the lens of Conv-BDI, this means designing a conversational agent with the capability to work towards an increasingly diverse range of Desires.

3.3 Intentions

The last component from the classical BDI model is Intentions. Intentions should not be confused with intentionality[1], which we view as the capability of the agent to decide upon and commit to long-term goals in the context of its Purpose. An Intention is a Desire that has been committed to by the system, otherwise seen as a "Desire in Focus." As Rao et al. (1995) describe it: "[T]he intentions of the system capture the deliberative component of the system." The Intention is chosen based on the Desire the system most immediately needs to address for the user. For example, if the user wishes to book a train and a hotel room, the system will have two Desires: Book the user a train, and book the user a hotel room. In strictly BDI terms, the system seeks to act such that in the *future* world state (the Beliefs), the user has a train and a hotel reservation. Of the two, it must choose one or the other to accomplish before proceeding to the second. Because the Intention of the system (and the user's own intentions) may change, there is a need to keep track of the current Intention with respect to the state of the system's Beliefs.

Conventionally, an Intention is grouped with a discrete set of actions that work towards fulfilling it. With respect to conversational agents, each dialogue act is viewable as an action in itself. Other actions such as gestures or movement in embodied agents are also actions a system might take. The system may also employ other actions for the specific purpose of belief state updates, e.g. information retrieval with API calls. While systems based on response templates may have a relatively limited number of actions per Intention, freeform generation from LLMs allows a substantially larger set of abstract actions to be taken in pursuit of an Intention.

4 Extending BDI: Purpose, Guidelines and Actions

In this section, we define three further elements as additional components to the core BDI model as we described previously. These elements extend BDI to describe a layer of high-level system design necessary for an effective conversational agent. These elements are *Purpose*, *Guidelines*, and *Actions*. Within Conv-BDI, the Purpose of a conversational agent provides the "why" of the system that is necessary to define the scope of its Desires and Intentions. Guidelines specify constraints upon the system's actions within the scope of the purpose. Meanwhile, system Actions are given an extended description beyond their role as means to complete a system Intention as in many earlier descriptions achitectures based on BDI.

4.1 Purpose

Beyond the initial attributes described in the BDI model, the task of designing a conversational agent implies further considerations. While many

[1]Our usage of intentionality is also distinct from usage in philosophy, cf. https://plato.stanford.edu/entries/intentionality/

conversational agents are flexible and capable of handling numerous scenarios or domains, any practical conversational agent will have an intended scope of use by design. In this sense, we consider that an additional element of a model should reflect this design consideration and represent the core *reasons* for the system's existence. The *Purpose* of the system is thus the conceptual starting point for the model. The Purpose is given by the developers to define the scope of what the system should accomplish in general. In this sense, the first element of the Conv-BDI model is the Purpose, and the other elements are defined in relation to an initial broad definition of the model's design. The Desires in scope for the model are defined by the system's Purpose. In simple terms, this means that any Desires that are not congruent with the Purpose are defined as out of scope. Any Desires that are congruent with the system Purpose form the set of all possible Desires for that system. As in the classic definition of the BDI model, the Intention of the system is the selected Desire that is to be worked towards.

4.2 Guidelines

The second additional element we add are *Behavioral Guidelines*. The system's Guidelines are informed by the Purpose, and are selected to define the bounds of *how* the system should interact with a user in carrying out its Purpose. For instance, a system whose purpose is to provide travel recommendations to a user might have guidelines as basic as "be friendly" and "be concise" while also having more specific guidelines like "make the user excited". These are general statements in a similar fashion to what is commonly included in the prompt of an LLM in many systems, and can be practically implemented in the same way. Additionally, Guidelines may also apply to the style or formatting of non-linguistic actions. For example, a system may be asked to return JSON formatted output or a string tailored to specific API calls, e.g. Dialport (Zhao et al., 2016). In general, Behavioral Guidelines as a component of the system provide an outline for how the system actions and reactions should be performed, separate to the relation of the action to a goal or Intention.

4.3 Actions

Detailed descriptions of the role of *actions* taken by the system are not always given specific attention within models of agents in the BDI framework. In Conv-BDI, we consider actions as operating in a dual role of both expressing the system's Intention and updating its Beliefs. The core relevance of this aspect is that as actions taken by the system are accounted for in its Beliefs, they subsequently affect the system's future Desires and Intentions. In addition, we give additional attention to actions in order to characterize them with respect to the comparatively large and complex space of dialogue actions available to modern LLM-based systems.

Every Intention may be associated with actions that work towards achieving it. In the classical setup of a BDI agent, such actions are defined as a discrete set of formal, logical units or steps that should be taken to achieve the goal expressed by the Intention. However, in contemporary work relying on LLM components, a strict mapping of intentions to actions or a plan library containing a fixed set of discrete actions is no longer necessary nor even desirable. Actions in the form of dialogue acts and utterances in general need not be classified according to a specific logical form or dialogue act, but nonetheless may still be usefully conceived of as discrete logical units. In non-verbal modalities such as e.g. API calls to model-external components or systems, some structured representation of the action is necessary. Insofar as actions should be defined as discrete operations, they are most usefully framed with respect to achieving one or more Intentions, as in the classic BDI setup.

Actions can be mutually exclusive with each other, and can also work towards multiple Intentions. Different actions can also work to achieve the same Intention, and one might take multiple paths towards the same goal. Actions can be viewed in terms of Reinforcement learning. Viewed in terms of reinforcement learning, the difference between taking two alternate sequences of actions to the same goal may result in a different reward. In this sense, task success by the system is the completion of Intentions expressed as a function. Reward also relates to the Guidelines of the system, for instance a guideline of "be concise" implies a shorter sequence of actions yielding higher reward, all else equal. However, expressed as language, Guidelines are not a formal mathematical definition of reward.

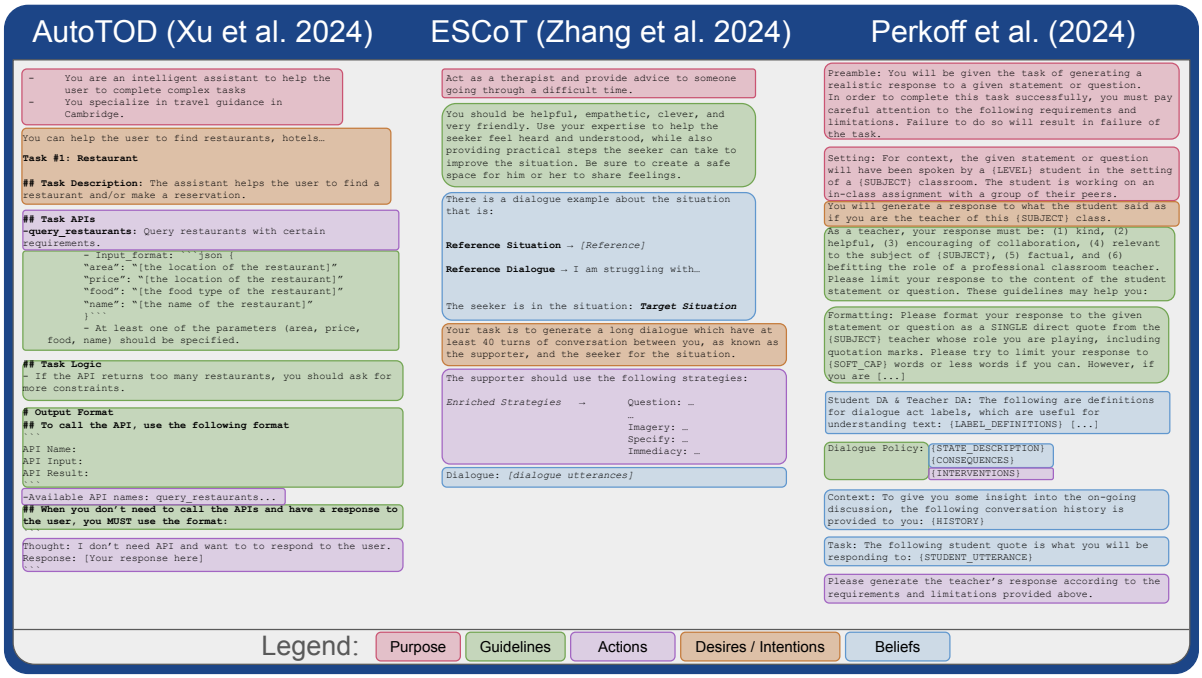

Figure 3: Prompts from Xu et al. (2024), Zhang et al. (2024), and Perkoff et al. (2024) with slight modifications to their formatting for presentation, highlighting the elements of the prompt structures as Conv-BDI elements. Note that Beliefs are not highlighted in the first model due to a specific instance of the database query results (the Beliefs in the prompt) not being included in the format. In the third model, the prompt is structured as a template with specific instances of domains and context inserted later. Within these models, each of the elements is represented in the prompt structure. However, it is not strictly necessary to do so. Modular approaches may dedicate modules to these functions and thus provide implicit rather than explicit signals to the LLM.

5 Conversational Agents with LLMs in the Lens of Conv-BDI

To concretely illustrate Conv-BDI in practice, we look at how it is realized in contemporary conversational agents, most often using LLMs. LLMs are now often a core technology for implementing conversational agents, and frequently rely on prompting to control system output, while information external to the model's parameters may be integrated with techniques such as RAG.

5.1 Prompting & Verbalization

With regard to LLM input, the Conv-BDI elements may be expressed as natural language within the prompt added to the LLM's context. An LLM prompt can be decomposed into chunks relating to several of these elements. For instance, a relatively fixed part of the prompt will be derived from the Purpose and Guidelines, although which Guidelines are applicable in a given term may change. Similarly, elements in the system Beliefs may change or become relevant at each dialogue turn. The system's Purpose and Guidelines can be expressed as direct instructions to the agent, defining the role it plays and what output should look like. Typically, the Purpose of the system is expressed as part of a "preamble" in the prompt defining what the system's role should be (Zamfirescu-Pereira et al., 2023).

The system Beliefs specifically may be partially expressed as a verbalization of any structured or background knowledge (for instance, a knowledge graph). Such verbalizations represent which part of the system's knowledge is in focus for the LLM's use, but does not necessarily represent the entirety of the knowledge available to the system. For instance, a large knowledge base such as Wikipedia may underlie the LLM's responses and be queried as needed as part of the system's Beliefs. Along with the system's Purpose, the Beliefs are the basis for the system's Desires. In concrete terms, this may be expressed in recent models elements such as Chain-of-Thought reasoning or "thinking" tokens e.g. (Guo et al., 2025), whereby the system makes use of the existing knowledge to sort through how this information should be processed within its parameters.

5.2 Control Signals & API Calls

Where traditional approaches to BDI agent design include a Plan Library of fixed plans to carry out system Intentions, we view the control of a conversational agent as effected by what can be described as control signals (Wagner and Ultes, 2024). At a broad level, a control signal to an LLM is a dialogue action to be taken by the system, explicitly expressed in order to guide the output of the model. Another example would be the support strategies in emotional support conversations as described by Liu et al. (2021). Likewise, Zhang et al. (2024) made use of such strategies for an emotional support agent, defined in the model prompt as shown in Figure 3. The strategies described in that work are actions that work towards a set of "stages" in the process of assisting the user.

Where the Purpose of ESCoT is to provide the user emotional support, the system sequentially acts to complete the three stages in the design. Each stage can be viewed as a Desire, which are individually taken as Intentions in turn by the system. The system then performs the associated actions in dialogue to fulfill them. Thus, viewed within the Conv-BDI framework, the control signal at a turn t is created from an action a associated with an intention $\mathcal{I}$ as well as the task data pulled from the graph (i.e. the belief state $\mathcal{B}$), which can be verbally represented in the prompt. Based on the Intention $\mathcal{I}$ the system is working to achieve, the system chooses an action and relevant knowledge from $\mathcal{B}$ to create a signal for the LLM generator.

6 Existing Approaches in the Conv-BDI Framework

We now consider how existing approaches to conversational agent design can be usefully characterized using the Conv-BDI framework. While numerous architectures including handcrafted, modular, or purely LLM-based are used to realize conversational agents, Conv-BDI gives an abstract characterization of the conceptual parts that is applicable across these architectures. Each of the components of Conv-BDI can be realized in diverse ways, either as specific modules, elements of a prompt in an LLM, or implicitly as part of the system's architecture. To demonstrate Conv-BDI as a conceptual framework describing conversational agent design in the NLP literature, we look at several models for comparison. These models are Conv-BDI within both the earlier Hidden Information State (HIS) model of Young (2006), the finite-state based Iris model of Fast et al. (2018) and more recent models such as the MOSS model of Liang et al. (2020) and the AutoTOD model of Xu et al. (2024).

6.1 Hidden Information State (Young, 2006)

As a POMDP-based system, the HIS model of Young (2006) bases the conversational agent's policy on a belief state representing the system's partial observations of the world state (that is, accounting for uncertainty in its observations). As described in the original paper, the HIS model is a task-oriented conversational agent designed to assist users in specific domains. As it predates LLM models where a prompt explicitly describes the role of the agent, the Purpose of this model is implicit in its design (that is, help the users within its domain). The POMDP model makes use of a sophisticated approach to belief state estimation, ultimately serving as input to the policy module. The HIS belief state directly corresponds to the Beliefs of Conv-BDI.

Moving further, the HIS model takes actions with respect to *user* goals, which is not strictly the same as the agent's goals. However, the Desires and Intentions of the HIS model can be taken to be the accomplishment of the user goals, which in the HIS model are subdivided into "equivalence classes". These classes describe states wherein at a given time t, states of the same class share the same next action to achieve their goals . Based on these equivalence classes, the belief state of the model can also be refined using ontological rules that partition the belief state but do not update it.

6.2 Iris (Fast et al., 2018)

Next, we examine how Conv-BDI can describe a conversational agent using a handcrafted policy. The Iris model of Fast et al. (2018) uses a finite-state model for dialogue state tracking and its dialogue policy. In such a model, the possible dialogue states in a conversation are modelled as a finite sequence of steps, where the possible transitions between the dialogue states are predefined. Transitioning from one dialogue state to another is associated with an action on the part of the agent in response to user input.

In handcrafted models, the Purpose of the model is inherent to the architecture, as the scope of the system's outputs are manually defined by

Explicitly Defined Components in Conversational Agents

Paper	Type	Purpose	Guidelines	Beliefs	Desires/Intentions
Young (2006)	*Task-Oriented*	✗	✗	✓	✗
Fast et al. (2018)	*Task-Oriented*	✗	✗	✓	✗
Liang et al. (2020)	*Task-Oriented*	✗	✓	✓	✓
Xu et al. (2024)	*Task-Oriented*	✓	✓	✓	✓
Perkoff et al. (2024)	*Task-Oriented*	✓	✓	✓	✓
Roller et al. (2021)	*Open-Domain*	✗	✗	✓	✗
Bae et al. (2022)	*Open-Domain*	✓	✓	✓	✓

Table 1: Comparison of different conversational system architectures in terms of Conv-BDI components, describing whether the component is implemented as a module, prompt element or otherwise represented (partially or fully) as an *explicit* part of the model (marked here with a check mark ✓), or completely *implicit* in the architecture by design, e.g. within neural network parameters or the model states as by Young (2006).

the designer. Likewise, the Guidelines are expressed by how the states connect to each other. That is: what actions should be executed in which context. The system Desires are defined by tasks in scope of the system's Purpose, specifically Data Science tasks in Iris. The Intention is then the task the system is currently working on for the user through conversation. Lastly, composition of system functions in Iris is enabled by saved information passed between states as a dictionary. This information corresponds to the Beliefs of the system in the Conv-BDI framework.

6.3 MOSS (Liang et al., 2020)

As a more recent example, we also observe the MOSS model (Liang et al., 2020). MOSS is a modular approach relying on a single encoder used by a number of different decoders for language understanding and dialogue policy. This system makes use of both a belief state estimate based on the dialogue history (optionally through an NLU component along with a DST module) and queries to a database. This system was demonstrated for restaurant recommendations, and is thus designed as a task-oriented conversational agent. The Purpose of the model is to suggest appropriate restaurants to the user, with the Guidelines on its behavior being comparatively limited to the constraints provided by the users themselves. The authors of this work present the option (though not necessity) of a dialogue policy learning module which predicts explicit logical representations of the system's actions.

6.4 AutoTOD (Xu et al., 2024)

We also observe the AutoTOD model, which is a non-modularized conversational agent design (Xu et al., 2024). AutoTOD contrasts with POMDPs in being based on an LLM component, with the system relying solely upon prompting strategies to direct the conversational agent. Nonetheless, this system also may be broken down into Conv-BDI elements. As shown in Figure 3, the prompt can be subdivided into sections of text providing the individual Conv-BDI elements. The scenario description provides the Purpose of the agent explicitly, contrasting with the implicit purpose in the HIS model. Within this purpose, the designers include a description of the tasks the system might handle, in the figure specifically the task of finding a restaurant. This corresponds to the Desires and Intentions of the system, that is, objectives for the system to select and then work towards through dialogue and API calls. This system prompt also include a number of guidelines for the model's responses and output, for example that at least one parameter on the restaurant selection should be specified in the API calls. Besides dialogue responses to the user, the possible system actions are explicitly provided in the list of API calls.

6.5 Other Models

Lastly, we consider Conv-BDI components in several additional models that demonstrate the diversity of contexts in which they can be employed. The model presented by Perkoff et al. (2024) elicited appropriate teacher-like responses for an educational conversational agent by inserting constraints into the prompt. The model relies on ex-

plicit extraction of dialogue state (Beliefs) as well as specifically enumerated dialogue acts for the agent (Actions), both of which are included directly in the system prompt. As shown in Figure 3, the model prompt also includes a preamble elucidating the system's Purpose, Desires that are tailored to match the subject of the dialogue, and Guidelines constraining the manner of the response generation.

To also describe *open-domain* conversational agents with Conv-BDI, we also compare the Generative BST model (Roller et al., 2021) and the model of Bae et al. (2022). As shown in Table 1, the Generative BST model lacks explicit representation of several components. As a sequence-to-sequence model, it does not incorporate an explicit expression of Purpose and Guidelines in the sense that previous model prompts exhibit. Rather, the model is endowed with these elements along with the Desires and Intentions implicitly within the model parameters through the training process. For this reason, adaptation of these components requires retraining or fine-tuning on new data, in contrast to prompt adaptation in other models.

By contrast, the model of Bae et al. (2022) explicitly defines each of the Conv-BDI components. In particular, they design the open-domain system with role specification that includes the system's Purpose and Guidelines. The role specifications in their system include constraints upon politeness and out of scope utterance categories. Simultaneously, the system's Desires and Intentions are framed in terms of initiating conversation and conversing over general topics.

7 Conclusion

This paper presents the Conv-BDI framework for conversational agents, a new conceptual model of the elements needed to build conversational agents in the context of contemporary technological advances. With the BDI model for autonomous agents as a basis, we identify two further elements that contribute to conversational agent design: Purpose and Behavioral Guidelines. We additionally elaborate on the role of actions in this extended model. As a general-purpose and abstract framework, a conversational agent may be implemented within the scope of Conv-BDI in different domains and architectures. The Conv-BDI components we have described characterize the design of contemporary conversational agents spanning open-domain and task-oriented systems as well as modular and end-to-end architectures.

Limitations

This paper investigates a theoretical perspective of conversational agent design from the perspective of the BDI model. Formal definitions of the components of BDI are not given here, though they may be found in the original sources defining it. This paper also observes a selected number of models from the NLP literature to illustrate and justify the Conv-BDI framework, however numerous other models for conversational agents exist and may warrant analysis as well.

It should also be noted that the Beliefs, Desires, and Intentions of the BDI model describe specific characteristics with respect to an artificial agent and should not be confused with the understanding of such terms in psychology. Nonetheless, the similarities to human psychology or lack thereof within artificial conversational agents may also be a worthwhile topic for analysis and comparison.

Further, empirical study of conversational agents would be a valuable and necessary addition to this line of inquiry. Illustrating Conv-BDI with experiments to demonstrate the effect of different Purposes, Guidelines, or BDI components in live settings would help elucidate the utility of this model.

References

Carole Adam and Benoit Gaudou. 2016. Bdi agents in social simulations: a survey. *The Knowledge Engineering Review*, 31(3):207–238.

Atheer Algherairy and Moataz Ahmed. 2024. A review of dialogue systems: current trends and future directions. *Neural Computing and Applications*, 36(12):6325–6351.

Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee, and Woomyoung Park. 2022. Building a role specified open-domain dialogue system leveraging large-scale language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2128–2150, Seattle, United States. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv:2108.07258*.

Matteo Bortoletto, Lei Shi, and Andreas Bulling. 2024. Neural reasoning about agents' goals, preferences, and actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 456–464.

Jan Broersen, Mehdi Dastani, and Leendert van der Torre. 2005. Beliefs, obligations, intentions, and desires as components in an agent architecture. *International Journal of Intelligent Systems*, 20(9):893–919.

Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. 2022. A literature survey of recent advances in chatbots. *Information*, 13(1):41.

Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjalmsson, Hao Yan, et al. 2000. Human conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents*, pages 29–63.

Kushal Chawla, Hannah Rashkin, Gaurav Singh Tomar, and David Reitter. 2024. Investigating content planning for navigating trade-offs in knowledge-grounded dialogue. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2316–2335, St. Julian's, Malta. Association for Computational Linguistics.

Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S Bernstein. 2018. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.

David Griol, Lluís F Hurtado, Encarna Segarra, and Emilio Sanchis. 2008. A statistical approach to spoken dialog systems design and evaluation. *Speech Communication*, 50(8-9):666–682.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tom Holvoet and Paul Valckenaers. 2006. Beliefs, desires and intentions through the environment. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 1052–1054.

Alexandre Yukio Ichida and Felipe Meneguzzi. 2023. Modeling a conversational agent using bdi framework. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 856–863.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Weixin Liang, Youzhi Tian, Chengcai Chen, and Zhou Yu. 2020. Moss: End-to-end dialog system framework with modular supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8327–8335.

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2024. Decision-oriented dialogue for human-ai collaboration. *Transactions of the Association for Computational Linguistics*, 12:892–911.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Michael McTear. 2021. Rule-based dialogue systems: Architecture, methods, and tools. In *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*, pages 43–70. Springer.

E. Margaret Perkoff, Angela Maria Ramirez, Sean von Bayern, Marilyn Walker, and James Martin. 2024. "keep up the good work!": Using constraints in zero shot prompting to generate supportive teacher responses. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–138, Kyoto, Japan. Association for Computational Linguistics.

Anand S Rao and Michael P Georgeff. 1997. Modeling rational agents within a bdi-architecture. *Readings in agents*, pages 317–328.

Anand S Rao, Michael P Georgeff, et al. 1995. Bdi agents: from theory to practice. In *Icmas*, volume 95, pages 312–319.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Zachary P Rosen and Rick Dale. 2024. Llms don't "do things with words" but their lack of illocution can inform the study of human discourse. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Pearson.

Yanet Sánchez, Teresa Coma, Antonio Aguelo, and Eva Cerezo. 2019. Abc-ebdi: An affective framework for bdi agents. *Cognitive Systems Research*, 58:195–216.

Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. 2023. Practices for governing agentic ai systems. *Research Paper, OpenAI*.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19:10–26.

Stefan Ultes, Lina M Rojas Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Inigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gasic, et al. 2017. Pydial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78.

Antonio Vetrò, Antonio Santangelo, Elena Beretta, and Juan Carlos De Martin. 2019. Ai: from rational agents to socially responsible agents. *Digital policy, regulation and governance*, 21(3):291–304.

Nicolas Wagner and Stefan Ultes. 2024. On the controllability of large language models for dialogue interaction. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–221, Kyoto, Japan. Association for Computational Linguistics.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and Heyan Huang. 2024. Rethinking task-oriented dialogue systems: From complex modularity to zero-shot autonomous agent. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2748–2763, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Steve Young. 2006. Using pomdps for dialog management. In *2006 IEEE Spoken Language Technology Workshop*, pages 8–13.

JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2016. Dialport: Connecting the spoken dialog research community to real user data. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 83–90. IEEE.

Yangyang Zhao, Ben Niu, Mehdi Dastani, and Shihan Wang. 2024. Bootstrapped policy learning for task-oriented dialogue through goal shaping. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4566–4580, Miami, Florida, USA. Association for Computational Linguistics.