Advances in Cognitive Systems 5 (2017) 1–12 Submitted 3/2017; published 5/2017

# A Cognitive Systems Analysis of Personality and Conversational Style

**Pat Langley** PATRICK.W.LANGLEY@GMAIL.COM
Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306 USA

## Abstract

In this paper, we analyze personality from a cognitive systems perspective. We review briefly some popular theories of this phenomenon, arguing that they offer descriptive rather than deeper accounts. Next we note four core aspects of personality that deserve explanation and propose an alternative theory that claims individual differences result primarily from high-level cognitive structures and processes. We elaborate on this theory by providing details about representations and the mechanisms that operate over them, illustrating the ideas with examples related to conversational style, which is often associated with personality. We conclude by discussing related work and noting directions for further research in this important and challenging area.

## 1. Introduction

A key factor that influences human behavior is *personality*. Although an individual's environmental situation and domain knowledge undoubtedly determine the choices available to him, personality dominates which options he pursues. A complete theory of the mind would include an explanation of this influence in terms of the structures and processes that underlie personality. Such an account would have not only scientific interest but also useful applications, as it would support more compelling synthetic characters for interactive entertainment, more distinctive robots for human-machine teams, and more enjoyable conversational interfaces for mobile telephones.

There has been some AI research on personality, but most of the work has been limited in scope and, we argue, descriptive of the known phenomena rather than explanatory. In this paper, we champion two nonstandard positions on this topic. First, we claim that a central aspect of personality involves interaction with other agents. This is exhibited most directly in conversational style, which leads us to focus here primarily on dialogue. Second, we posit that personality is a high-level cognitive phenomenon that is best explained in terms of abstract structures and processes. Both positions make our work highly relevant to the cognitive systems movement, as will become apparent later. Our aim is to provide a broad account of personality and its relation to other facets of intelligence, as opposed to fitting results from specific experiments.

In the sections that follow, we review some influential psychological theories of personality, identify the phenomena we hope to explain, and state the main tenets of our theory. After this, we provide details about the cognitive structures and processes that we hypothesize underlie personality, drawing examples from conversational style. In closing, we discuss related work in the area and then note some promising topics for additional research.

## 2. Mainstream Accounts of Personality

The study of personality has been a major theme in psychology and has produced a number of alternative theories (Ewen, 2009). We do not have the space to review them fully, so we will focus on two popular accounts which take different positions on key issues that are especially relevant to our later discussions of this intriguing topic. Many other theoretical frameworks are linked directly to psychotherapy, which is not our focus here, even some that incorporate ideas from cognitive psychology (e.g., Kelly, 1955).

Some theories attempt to explain personality in terms of *behaviorist psychology* (Skinner, 1969), which claims that behavior in humans and animals results from stored stimulus-response pairs. This framework downplays, and in extreme variants denies, the role of mental structures and processes, emphasizing instead direct connections between perception and action. Behaviorism also gives a central role to learning of these connections from reward through mechanisms of instrumental conditioning. The implication is that differences in people's personalities arise from different sets of stimulus-response links that are themselves learned from their experiences. This in turn suggests that personality is highly malleable and subject to at least gradual change over time.

In contrast, *trait theories* of personality posit a set of attributes or dimensions along which people differ. These typically assume that each person has fixed values for these traits, which in turn influence their behavior. For example, Digman's (1990) theory of personality proposes five high-level traits: *openness* – appreciation of new and varied experiences; *conscientiousness* – exhibiting self discipline and planned behavior; *extraversion* – stimulation from others' presence; *agreeableness* – compassion for and cooperation with others; and *neuroticism* – experience of unpleasant emotions. These traits appear to describe personality differences that arise in many cultures, but there are a number of variants, with others positing up to 16 distinct dimensions (Cattell, 1957).

Trait theories have been adopted in most AI research on synthetic characters (e.g., Rosseau & Hayes-Roth, 1997), as they offer a simple means to instill behavioral differences among agents. However, although the framework offers a useful *description* of ways in which human personality can vary, it does not offer an *explanation* in terms of structures and processes. Thus, like behaviorist accounts, they make no contact with results on high-level cognition, although for different reasons. A few theorists, like Ford (1992) and Mischel (2004), have studied personality from a cognitive perspective, but their work has not drawn the same attention. We elaborate on some of their ideas in our own treatment of the topic.

## 3. Behavioral Phenomena and Theoretical Claims

Science is concerned not merely with observations and not solely with theories, but with their relationship. This means that, in our analysis of personality, we must both identify the main phenomena that we associate with this term and propose a set of principles that account for these regularities. We will focus here on four main personality-related phenomena:

- *Personality differs across people*. Just as humans differ in their height, hair, and other physical features, they also differ in their behavioral styles.
- *Personality is stable over time*. A person's distinctive behavioral style is reasonably fixed and, if it changes at all, shifts slowly.

*Table 1.* Some aspects of personality viewed as important enough to merit words in English.

| | | | |
|---|---|---|---|
| Friendly | Distant | Organized | Careless |
| Caring | Unconcerned | Thoughtful | Thoughtless |
| Selfless | Selfish | Giving | Greedy |
| Persistent | Relenting | Stubborn | Compromising |
| Judgmental | Forgiving | Relaxed | Tense |
| Loyal | Disloyal | Reliable | Unreliable |
| Trusting | Suspicious | Confident | Timid |
| Brave | Cowardly | Open minded | Dogmatic |

- *Personality has global influence on behavior*. These stable regularities cut across many classes of situations and areas of expertise; they are largely domain independent.
- *Personality has both coarse-grained and fine-grained aspects*. Although many behavioral regularities are high level, some *idiosyncratic* facets or *quirks* involve low-level behavior.

These regularities are widely recognized and, indeed, there almost serve as the definition of what we mean by the term 'personality'. As a result, they are seldom mentioned in the psychological literature on the topic, which instead focuses on detailed studies, but this makes them natural targets and we will adopt them as the main phenomena to be explained in this paper. Other results certainly deserve attention, but these provide a good starting point for our analysis.

Table 1 presents some common English words that refer to personality characteristics. This list is very incomplete, but each term describes some aspect of behavior that differs across people, that is typically stable over time, and that has global rather than domain-specific effects. These generic terms do not include any idiosyncratic behavioral tendencies, but there is little doubt that personality sometimes involves fine-grained regularities (e.g., a refusal to shake hands) in addition to coarse-grained ones. Note also that many of these terms deal with interpersonal interactions, which suggests that recurring patterns of social behavior constitute an important element of personality. There are certainly facets of personality not directly related to social behavior, as reflected by words like 'organized' and 'energetic', but this does not reduce its relevance to interaction.

We desire a computational account for these regularities, but what form should it take? We have already argued that trait theories offer shallow descriptions rather than deep explanations, but there are different ways to approach the topic within the cognitive systems paradigm. For example, we might claim that personality is linked closely to the cognitive architecture, which is typically viewed as stable and which has global effects on behavior. Traits like *confidence*, *persistence*, and *bravery* could map onto architectural parameters that differ across people but do not change over time, which suggests they may be innate. However, it is less clear how other terms from Table 1 might be handled by such an account. Words like *friendly*, *selfish*, *compromising*, and *judgmental* deal primarily with social contexts and have no obvious connection to parameters that would arise in a theory of the cognitive architecture.

Instead, we propose here an alternative theory of personality that takes a cognitive systems perspective and emphasizes the central role of *knowledge*. This account revolves around four separate but complementary postulates:

- *Personality is determined by long-term cognitive structures*. More specifically, personality-related content is stored as rule-like elements that generate goals and tasks, with different people having either different mental structures or different priorities on them.
- *These cognitive structures are primarily general and domain independent*. In particular, the elements specify abstract relations among beliefs, goals, and relationships, typically making no reference to domain-level predicates.
- *Although domain independent, these structures occur at multiple levels of specificity*. Despite their abstract character, personality-related elements can sometimes refer to narrow (idiosyncratic) elicitation conditions and to detailed, concrete activities.
- *Personality structures exert a metacognitive influence on thinking and action*. In other words, these elements are responsible for generating the top-level goals and tasks that a person pursues, as well as modulating problem solving and execution to achieve them.

The first three principles concern representational stances, whereas the final assumption deals with issues of cognitive processing. We elaborate on the former in the next section, after which we turn to the fourth postulate. We will illustrate the theory with examples related to conversational styles, as they provide useful intuitions, although we maintain that it applies to other forms of interaction besides dialogue, and even to nonsocial settings where a person operates in isolation.

Note that our theoretical assumptions are not truly inconsistent with trait theory. Abstract rules for generating goals and tasks might well be viewed as corresponding to values for particular dimensions such as extroversion. However, we argue that they account for the primary phenomena at a deeper level, in terms of cognitive structures and processes. This approach to personality is not entirely new. As we discuss later, both Rizzo et al. (1999) and Evans (2011) have developed computational models that incorporate some of these ideas. However, they did not examine their influence on topics like conversational style, the aspect of behavior that we emphasize here, or link them to notions of metacognitive control.

## 4. The Cognitive Structure of Personality

Before we can discuss the processes responsible for personality, we must first examine the mental structures over which these mechanisms operate. Our role model is Ortony, Clore, and Collins' (1988) analysis of the cognitive structure of emotions, which offered a detailed representational framework for another area that has received relatively little attention in AI and cognitive science. We start by proposing some general constraints on personality structures and then clarify them with examples related to conversational style.

### 4.1 General Structural Characteristics

As mentioned earlier, the first postulate of our theory is that personality is determined by structures stored in long-term memory. In our framework, differences in behavioral style result not from parameters associated with the cognitive architecture but rather from differences in the *knowledge*

that it contains. The presence of this content in long-term memory means that it changes slowly, if at all, and that it has strong structural and relational components, although there may also be numeric annotations to these structures.

Moreover, there is little question that many factors contribute to a given personality. As we have noted, trait theories vary in the number of dimensions they assume, with some positing as many as 16 traits and others only five, but they agree on the need for multiple elements. Within our framework, this translates into the need for multiple elements in long-term memory that are both modular and compositional in character, as we must state them separately but combine them as necessary. Each element corresponds to a different aspect of behavioral style, with their combination making up the total personality of the agent. This suggests that we encode the knowledge as rules or other notation that exhibits these features.

We will also argue that these cognitive structures must be relational in nature. This follows partly because their global influence on behavior requires them to be abstract and, typically, to eschew domain-level predicates. When we say that someone is organized or persistent, we are referring to their behavior across many different contexts. Another reason is that, as noted earlier, many aspects of personality concern interactions with other agents, and thus necessitate encoding interpersonal relations. We often claim that someone is forgiving or stubborn in their interactions generally, not in only one type of situation. Moreover, we maintain that these structures describe the situations in which an agent creates goals or adopts tasks; this requires specifying relations between generalized conditions and generalized effects using rules or some analogous formalism.

Im summary, we postulate that three distinct forms of knowledge influence activity in general and conversational behavior in particular. These include skills that let the agent act in the environment, conceptual rules that let it draw inferences, and motives that lead to creation of goals. In the remainder of the section, we discuss each of these knowledge categories in turn. We hold that personality resides mainly in the third class of cognitive structures, motives, but that the other two, skills and concepts, are also required to enable it.

### 4.2 Skills and Conversational Style

We have characterized personality as related to an agent's *behavior*, but to carry out different activities, the agent must have knowledge that describes possible actions. Following Langley, Choi, and Rogers (2009), we will refer to these cognitive structures as *skills*. In their architectural framework, skills are organized in a hierarchy, in a manner similar to that in hierarchical task networks. Here we will assume, for the sake of simplicity, that skills occur at only one level and that they describe the effects of primitive actions under given conditions, much like the STRIPS and PDDL notations for operators used in many AI planning systems.

As noted earlier, conversational style is a natural area in which to illustrate our cognitive theory of personality. There has already been substantial research on the formal representation of speech acts, like making a statement in dialogue, asking a question, or rejecting a proposal, one of the earliest being Perrault and Allen's (1980) analysis. They proposed operator-like structures for some common conversational actions in terms of beliefs that serve as conditions for application, goals they aim to achieve, and effects they produce. Gabaldon, Langley, and Meadows (2014) have offered a similar analysis that they incorporate into a high-level dialogue system for procedural assistance.

Both treatments are interesting and relevant because they assume that conversational skills are general and domain independent. For instance, we might state a structure for informing another person about something in a notation similar to that for a STRIPS operator:

*inform(S, L, C)*
*conditions:* *believes(S, C), believes(S, not(believes(L, C)))*
*action:* **inform(S, L, C)*
*effects:* *believes(S, inform(S, L, C)), believes(S, believes(L, C))*

The conditions for this skill specify that the speaker (*S*) believes some content (*C*) and believes that the listener (*L*) does not believe it. In addition, the rule states the effects of the speech act, which include the speaker believing that he has informed *L* about *C* and his believing that *L* now believes this information. Variants of the *inform* skill are also possible. For instance, one can specify a *lie* speech act in which the speaker communicates content that he does not actually believe himself.

This structure refers to the domain content being conveyed, but it does not mention any particular domain predicates. Thus, it involves the same level of abstraction as our account of personality, which suggests that we may be able to use such rules to help account for differences in conversational style. We can write analogous structures for other types of speech acts, although the conditions and effects will differ. For instance, the skill for a *propose* action might lead S to believe L has acquired a goal that L did not have before their communication, rather than adopting a new belief. Alternatively, it might state that the listener only considers adopting it, with separate skills for *accept* or *reject* speech acts indicating the final decision.

To exhibit distinct personalities through conversational style, agents must use these various types of speech acts in different ways. For instance, consider five utterances that someone might use to encourage someone else to stop smoking in an elevator:

- Would you mind not smoking in the elevator?
- You know, smoking in the elevator isn't allowed.
- You really can't smoke in the elevator.
- You can't smoke here. Please put your cigarette out.
- Put that cigarette out now or I'll do it!

We associate these utterances with personalities that take quite different approaches to interaction. We view the first sentence as deferential, whereas the second is still polite but has a more authoritative flavor and the third statement takes an even firmer position. The fourth utterance makes the speaker's request still more explicit, and the final version verges on a threat if the listener does not comply. These variants have similar content but they convey very different personal styles.

Table 2 presents a few out of the many English adjectives that describe the manner in which we can present speech acts. Each row refers to one type of speech act, using categories that often appear in the literature. For instance, the statements in the elevator scenario are all examples of a *propose* speech act, which encourages the listener to adopt a goal of the speaker. As we have seen, a person can be polite, authoritative, or threatening when making a proposal. Similarly, we can be deferential, demanding, or impertinent when asking someone a question. English also uses some of these adjectives to describe aspects of personality, which does not seem accidental, as these characteristics are often reflected in a speaker's conversational style.

*Table 2.* Six categories of speech acts and adjectives that describe variations related to conversational style.

| Speech acts | Variations linked to conversational style | | |
|---|---|---|---|
| Inform | Complimentary | Insulting | Condescending |
| Propose | Polite | Authoritative | Threatening |
| Question | Deferential | Demanding | Impertinent |
| Acknowledge | Appreciative | Nonchalant | Flippant |
| Accept | Agreeable | Ingratiating | Insubordinate |
| Reject | Apologetic | Combative | Offended |

We can characterize such nuanced conversational actions with specialized versions of the skills associated with basic speech acts. Consider a variant of the *inform* rule that describes the conditions and effects for telling someone a fact the speaker might expect the listener to already believe:

*inform-expected(S, L, C)*
*conditions:* *believes(S, C), believes(S, not(believes(L, C)))*
*action:* **inform-expected(S, L, C)*
*effects:* *believes(S, inform(S, L, C)), believes(S, believes(L, C)), believes(S, believes(L, expected(S, believes(L, C))))*

This includes similar generic elements as a basic *inform* action, but it also contains an effect that suggests the listener should already have believed the content. Instilling such a high-level belief in the listener may have important indirect effects, such as eliciting intended emotions like embarrassment or guilt. Other variations on skills for speech acts would include different, but similarly abstract, application conditions and expected effects. We will see later how such beliefs can interact with other cognitive structures.

We should also note that a more complete account would include details about the manner in which the speaker delivers a given speech act. A conversant will typically communicate a compliment in a different tone of voice than he will an insult, and he will convey impatience or disdain at a different speed or different volume than he will other utterances. Our notation reflects this point by using distinct names in the *action* fields for, say, *inform* and *inform-expected*, but this merely acknowledges variations at the speech level and does not provide a detailed analysis of how one delivers the same content in different manners.

### 4.3 Goals and Motives

The specialized rules for varieties of speech acts described above provide the raw material to support differences in conversational style, but they do not offer reasons why some people use them and others do not, or why some use them in different situations. For this, we must turn to the speaker's goals, which we assume can influence his selection of which skills to execute. However, traditional treatments of goals hold they are similar to beliefs, in that they are concrete and relatively short term. Following Choi (2011) and Langley et al. (2016), we distinguish between concrete goals,

which reside in working memory, and generalized rules that specify the conditions for introducing them, which reside in long-term memory. To distinguish them from short-term goals, we will refer to these structures as *motives*[1] because they encode the underlying motivations for a person's behavior.

We assume that such motives take the form of rules that include a set of generalized conditions and a goal that should be activated when they are satisified. For instance, suppose that someone wants a person to be proud of him if he respects that person; we can state this motive as:

*wants(A, (believes(B, proud_of(B, A)))*
*conditions: believes(A, respects(A, B))*
*priority: 5.5*

Now suppose someone else has an 'eye for an eye' motive, so that if he believes someone has caused an event that disappointed him, then he desires to reciprocate in kind:

*wants(A, believes(B, disappointed(B, _)))*
*conditions: believes(A, disappointed(A, E)), believes(A, cause(B, E))*
*priority: 10.2*

This motive states that, under such conditions, one should create a goal for the other person to be disappointed about something as well, although it does not specify the details. These rules include a field that specifies the priority to associate with its created goal, although we might instead specify a numeric function that varies with the situation, as in Langley et al. (2016). Both rules describe an emotion the agent wants to instill in someone else, making them relevant for decisions about what speech acts to use in dialogues. We claim that such goal-generating rules underlie differences in conversational styles and in personality more generally.

A variation on this idea is that individual differences result not from divergence in the structures of goal-generating rules but rather in the priorities or weights associated with them. In this view, two people who behave very differently, and who exhibit distinct conversational styles, may have in their long-term memories exactly the same library of motives, but they may have quite different distributions of weights on them. We will return to this hypothesis later when we discuss cognitive processes that operate over these structures to produce personality.

### 4.4 Conceptual Knowledge

Both forms of knowledge that we have examined, skills and motives, can make reference to abstract predicates that an agent cannot observe directly and thus must infer from other sources. This means in turn that we need a third form of knowledge that lets the agent connect these conceptual predicates with lower-level ones that it can observe, at least in principle. We assume that these take the form of inferences rules which, like motives, are similar in structure to Prolog clauses in that they contain a head and a set of conditions or antecedents. The difference is that the heads of conceptual rules denote beliefs that the agent can infer, which are useful for description of situations, rather than goals it can generate, which are useful for prescription.

Naturally, many conceptual rules encode domain-specific knowledge about the environment (e.g., that secondhand smoke is harmful), artifacts (e.g., that elevators are enclosed spaces), and

1. Other researchers have used different terms for a similar idea. For instance, Talamadupula et al. (2010) label them as 'open-world quantified goals', which is considerably less pithy.

even behavioral norms (e.g., one should not smoke in elevators). These specify relations much like those utilized by traditional expert systems, which focus on domain-level inference and decision making. However, conceptual knowledge may also include content that refers to high-level, domain-independent predicates that describe relations among beliefs, goals, and other structures without referring to their domain-level content. These generic elements involve the same level of abstraction as many of the motives that we hypothesize underlie personality.

One important class of such structures concerns rules for *emotional concepts*, which often specify relations among an agent's goals, beliefs, and expectations. For example, we might state a rule for recognizing instances of the *disappointed* concept as

*disappointed(Agent, Event)*
  *conditions: wants(Agent, Event), expect(Agent, Event), belief(Agent, not(Event))* .

This includes a head that specifies the emotional predicate, the agent who experiences the emotion, and the target, in this case the description of a possible event. The conditions state that this emotion occurs when the agent has a goal for the event to occur, he expected that event to take place, but he believes it did not transpire. Similarly, we might specify a rule for the *jealous* concept as

*jealous(Agent, Other, Object)*
  *conditions: wants(Agent, possess(Agent, Object)),*
  *belief(Agent, not(possess(Agent, Object))),*
  *belief(Agent, possess(Other, Object))* .

This states that an agent is jealous of another agent if he wants an object, believes he does not possess it, and believes the other agent does have it.[2] These examples follow the tradition of Ortony et al. (1988), who specified abstract patterns associated with many familiar emotions. Such rules do not address the visceral aspects of emotional experience, but they suffice for our current purposes.

As we will see shortly, emotional concepts are important to personality and conversational style because the emotions they add to working memory can match against the conditions of motivational rules, which in turn generate the agent's goals. Thus, they serve as mediators between domain-level inferences about the environment and goals that drive behavior. However, we can also formulate specialized versions of these rules that produce emotional literals under alternative conditions, so that distinct agents become *proud*, *offended*, or *jealous* in quite different situations. To the extent that these differ across individuals in ways that affect behavior, they also constitute an important element of personality, as in Evans' (2011) treatment of this topic.

## 5. Cognitive Processes for Personality

Now that we have discussed the cognitive structures we claim underlie personality and conversational style, we can examine the hypothesized processes that operate over them. We assume an agent architecture that operates in successive cycles, much as in production systems (Klahr, Langley, & Neches, 1987), but with three distinct stages of processing. Each stage involves matching the conditions of rules against the contents of working memory, firing one or more of the matched rules, and carrying out its associated inferences and actions. Here we discuss each stage in turn,

2. The emphasis on belief here is essential. Someone can be disappointed or jealous about a situation that he believes is true even when the situation does not actually hold in the world.

covering them in the reverse order from the previous section. As before, we clarify the framework with examples related to dialogue, but the same processes can involve other forms of physical and mental actions to account for personality differences in other contexts.

### 5.1 Conceptual Inference

The first stage involves conceptual inference, which generates beliefs and other working memory elements from those already present. This process also takes into account newly added elements that describe input from the environment. For physical tasks, this input describes objects and events perceived in the environment; for conversational settings, it corresponds to utterances made by other agents. Simple versions of the module could assume that the basic speech-act types of these utterances (e.g., *inform*, *question*, *propose*, *reject*) are provided, along with their content, as in Gabaldon et al. (2014). More advanced versions could determine the speech act from its content and even judge its truthfulness.

This stage is responsible for applying domain knowledge to make domain-level inferences. In dialogue, these play a key role in driving conversation, as they provide new content that the speaker can communicate, as well as produce working memory elements that help him process utterances made by others. For instance, if you believe that someone is allergic to smoke, you can infer that its presence can cause medical problems. The architecture may apply this inference process not only to extend its own beliefs, but also to update its mental model of other agents' beliefs. During a dialogue, this can lead it to draw conclusions about the other participants' knowledge states, which in turn can inform its responses to their utterances.

Conceptual inference is also responsible for generating emotions. We have already described emotional concepts in terms of abstract, domain-independent relations among an agent's beliefs, goals, expectations, and attributions about others. The same process that gives domain-level inferences can also produce working-memory elements that specify an emotional predicate and its target. For example, if the speaker uses the *inform-expected* speech act to convey an expectation that the listener would know something, the latter can reasonably infer that the speaker is disappointed in him. Moreover, just as the architecture draws conclusions about others' beliefs based on their behavior or utterances, so can it use these mechanisms to make inferences about their emotions.

We will not take a position here on the details of this processing stage. The most straightforward approach to implement would carry out exhaustive deductive inference to generate the full closure of beliefs that are implied by available rules and the initial contents of working memory. However, one could also implement versions that use more limited and focused forms of inference, possibly guided by top-down factors like the relevance of consequents to active goals. There has also been work on the role of abductive inference in dialogue (e.g., Gabaldon et al., 2014), which introduces plausible default assumptions to explain the reasons for others' utterances. Regardless of its detailed operation, this stage updates beliefs based on inputs from the environment or others' utterances.

### 5.2 Goal Generation

The second stage involves goal production, which adds new goal elements to working memory and updates existing ones. As in conceptual inference, here the architecture compares the conditions of each motivational rule to the contents of working memory to determine which ones match. We

assume the system applies each of these rules in parallel, repeating the process until quiescence to handle motives that include goals in their condition side. These may include goals to instill in others not only beliefs about the world, but also to produce desired emotions. If agent *A* has an 'eye for an eye' motive described earlier and believes agent *B* has caused him disappointment, then *A* may adopt a goal to disappoint *B* in return. This stage can use motivational rules not only to generate its own goals, but to make informed guesses about the goals of other agents.

We assume that this stage not only adds goal structures to working memory, but also assigns them numeric priorities, as in Talamadupula et al. (2010) and Langley et al. (2016). This reflects the idea that two agents may have the same set of goals, but that they assign different relative importance to them, which in turn can lead to different behaviors. Moreover, the priority an agent assigns to a given goal may change over time, with its beliefs about the situation. We will not take a position here on the mechanism that produces such changes, but one candidate is a form of decay. Another involves making priorities a function of quantities associated with elements matched by conditions, which may include other goals. If so, then this processing stage would update the priorities for goals already in memory, which could shift their ordering in ways that alter the agent's behavior.

### 5.3 Activity Execution

Once the architecture has elaborated its mental state for the current cycle by drawing inferences and generating goals, it uses this information as context to bias its execution of activity. During this third stage, the system accesses each of its executable skills and determines which of them has conditions that match the current contents of working memory. As in the production-system paradigm (Klahr et al., 1987), we will refer to the matched rule instances as the *conflict set*. Thus, a conversational agent would consider its entire set of skills for speech acts. These differ in their conditions, so only some will match the current working memory, but more than one may be satisfied and, indeed, some may match in more than one way. For instance, if the inference stage has produced mixed emotions about another agent, then two distinct *inform* or *propose* rules could match, say with polite or rude overtones. In some cases, only one skill instance will apply, but this will be less common.

From this conflict set, the architecture selects one skill instance to execute. Traditional production systems use details about the matched conditions (e.g., recency) or the rules themselves (e.g., order) to make such a selection. In contrast, we claim that the agent takes into account how each skill instance's effects relate to the current goals. Briefly, the system examines which goals would be satisfied upon execution and sums their associated priorities, then determines which goals would be violated upon execution and subtracts their associated values. This total utility score reflects the tradeoffs among different goals affected by the candidate action.

Once the architecture has a score for each skill instance, it selects the member of the conflict set with the highest utility and executes it. In a conversational setting, this translates into selecting a specialized speech act that lets it achieve its highest priority goals while not clobbering others. Effectively, the architecture carries out one-step lookahead to determine the highest-utility action, much as in the performance element associated with many systems that incorporate reinforcement learning. The difference is that the utility function is factored into many different goals, only some of which are relevant to each skill instance. As a result, agents who generate different goals or that assign them different priorities will exhibit distinct personalities and conversational styles.

### 5.4 Metacognitive Aspects of Personality

Our goal-driven account of personality clarifies how such high-level mechanisms can influence not only an agent's physical behavior, such as a tendency to flee or fight, but also its cognitive processing, such as the amount of planning done before acting. The latter suggests that personality plays a *metacognitive* role that operates over and influences base-level cognition, as Cox (2007) has defined metacognition as mechanisms that inspect traces of cognition and modulate its operation. In our framework, motivational rules both inspect and alter the agent's goals, which are a primary source of control in cognitive systems.

Moreover, this connects our theory of personality to recent research on *goal reasoning* (Aha, Cox, & Muñoz-Avila, 2013), which also examines metacognitive issues. The key difference is that most work in this arena has focused on physical activity, whereas we have used these ideas to explain observed differences in people's behaviors, including their conversational styles. Emotions appear to play a similar modulating influence on cognition (Muramatsu & Hanoch, 2005), but the motivational rules that determine personality operate on an even higher level, since they can both match against, and create goals about, emotions themselves. Contrary to conventional accounts, personality arises from the highest levels of cognitive processing.

## 6. Discussion

At the outset, we enumerated four primary phenomena that we desired to explain. The first was that personality varies across people, who can exhibit quite different behavioral styles, especially in their interactions with others. Our theory accounts for this by positing different motivational rules that generate goals and priorities associated with them. The second was that personalities remain stable over time, in that they change slowly if at all. Our framework explains this fact by assuming that motives are stored as structures in long-term memory. The third phenomenon is that personality influences behavior consistently across many situations, which follows from the theory's tenet that many motives are encoded as abstract, relational rules that do not refer to domain predicates. However, personality can also produce fine-grained, idiosyncratic behaviors, which we explain by allowing highly specific motivational structures. Our theory appears to handle each of these key regularities. Moreover, it is not antithetical to traits or behaviorism: each motivation maps onto a single trait, whose priorities may change gradually due to learning. However, it offers a deeper account than either of these traditional frameworks.

We have borrowed ideas from a number of earlier efforts. One of our central tenets is that personality is linked to long-term cognitive structures. Rizzo et al. (1999) explored this idea in an extension to the Prodigy architecture that incorporated Ford's (1992) association of priorities with abstract goals, letting it encode personalities that produced different plans. Our theory replaces prioritized goals with conditional rules that generate them, building on work by Choi (2011) and Talamadupula et al. (2010). Evans (2011) has taken a similar approach to simulating personality differences in synthetic characters, although his rules map domain-level situations onto emotions, rather than defining them as abstract concepts. Our analysis of conversational style borrows the notion of abstract operators for speech acts (Allen & Perrault, 1980; Gabaldon et al., 2014), and our treatment of emotions adopts Ortony et al.'s (1988) claim that they involve abstract relations among

goals, beliefs, and expectations. Our theory builds on earlier ideas but combines them in novel ways to explain an understudied facet of intelligence.

However, the analysis reported here constitutes only the first step in an extended research programme. To demonstrate its viability, we should incorporate its assumptions into an implemented cognitive architecture and develop a number of agents with distinctive personalities. We should also extend the framework to include hierarchical skills, complement its reactive execution with planning mechanisms, and specify the details by which it generates speech acts during dialogues. To establish the theory's generality, we should demonstrate the resulting agents' behaviors on multiple conversational scenarios, and to show its breadth, we should encode a variety of familiar personality terms as motivational rules. Taking these additional steps will bring us far closer to a complete cognitive theory of personality and conversational style.

In summary, we reviewed four primary phenomena about personality: variation across people behavior in similar situations; stability over time; global influence on behavior; and both coarse and fine granularity. We used conversational style as a source of examples about the influence of personality on social interaction. In addition, we proposed a theory of these phenomena that posits central roles for cognitive structures and processes. The former comprise skills for specialized speech acts, conceptual rules that define predicates, including ones for emotions, and motivational rules that specify when to adopt goals and what priorities to assign them. The cognitive cycle that operates over these structures has three distinct stages: conceptual inference applies rules that update beliefs in working memory; goal generation invokes motives that add goals to working memory or alter their priorities; and execution finds skills that match working memory, selects one based on goal-related utility, and carries out the associated actions. Our analysis suggests that this theory accounts for the main features of personality noted earlier, including aspects of conversational style.

## Acknowledgements

This research was supported by Grant N00014-15-1-2517 from the Office of Naval Research, which is not responsible for its contents. We thank Richard Evans, Alfredo Gabaldon, Marc Pickett, Ted Selker, and Chris Tar for useful discussions that helped refine the ideas in this paper.

## References

Aha, D. A., Cox, M. T., & Munoz-Avila, H. (2013). (Eds.). *Goal reasoning: Papers from the ACS workshop*. Baltimore, MD: Cognitive Systems Foundation.

Allen, J. F., & Perrault, C. R. (1980). Analyzing intention in dialogues. *Artificial Intelligence*, *15*, 143–178.

Cattell, R. B. (1957). *Personality and motivation structure and measurement*. New York: World Book.

Choi, D. (2011). Reactive goal management in a cognitive architecture. *Cognitive Systems Research*, *12*, 293–308.

Cox, M. T. (2007.) Perpetual self-aware cognitive agents. *AI Magazine*, *28*, 32–45.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417–440.

Evans, R. (2011). Representing personality traits as conditionals. *Proceedings of the Fourth AISB Symposium on AI and Games* (pp. 35–42). York, UK.

Ewen, R. B. (2009). *An introduction to theories of personality* (7th ed.). Mahwah, NJ: Lawrence Erlbaum.

Ford, M. E. (1992). *Motivating humans: Goals, emotions, and personal agency beliefs*. Newbury Park, CA: Sage.

Gabaldon, A., Langley, P., & Meadows, B. (2014). Integrating meta-level and domain-level knowledge for task-oriented dialogue. *Advances in Cognitive Systems*, *3*, 201–219.

Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Norton.

Klahr, D., Langley, P., & Neches, R. (Eds.) (1987). *Production system models of learning and development*. Cambridge, MA: MIT Press.

Langley, P., Barley, M., Meadows, B., Choi, D., & Katz, E. P. (2016). Goals, utilities, and mental simulation in continuous planning. *Proceedings of the Fourth Annual Conference on Cognitive Systems*. Evanston, IL.

Langley, P., Choi, D., & Rogers, S. (2009). Acquisition of hierarchical reactive skills in a unified cognitive architecture. *Cognitive Systems Research*, *10*, 316–332.

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology*, *55*, 1–22.

Muramatsu, R., & Hanoch, Y. (2005). Emotions as a mechanism for boundedly rational agents: The fast and frugal way. *Journal of Economic Psychology*, *26*, 201âĂŞ-221.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. New York: Cambridge University Press.

Perrault, C. F., & Allen, J. F. (1980). A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics*, *6*, 167–182.

Rizzo, P., Veloso, M. M., Miceli, M., & Cesta, A. (1999). Goal-based personalities and social behaviors in believable agents. *Applied Articial Intelligence*, *13*, 239–271.

Rousseau, D., & Hayes-Roth, B. (1997). *Improvisational synthetic actors with flexible personalities* (Technical Report No. KSL-97-10). Knowledge Systems Laboratory, Department of Computer Science, Stanford University, Stanford, CA.

Skinner, B. F. (1969). *Contingencies of reinforcement: A theoretical analysis*. New York: Appleton-Century-Crofts.

Talamadupula, K., Benton, J., Schermerhorn, P., Kambhampati, S., & Scheutz, M. (2010). Integrating a closed world planner with an open world robot: A case study. *Proceedings of the Twenty Fourth AAAI Conference on Artificial Intelligence* (pp. 1561-1566). Atlanta: AAAI Press.