

Managing personality influences in dialogical agents

Jean-Paul Sansonnet¹, François Bouchet²

¹*LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France*

²*SMART Laboratory, McGill University, 3700 McTavish Street, Montreal, Canada
jps@limsi.fr, francois.bouchet@mcgill.ca*

Keywords: Cognitive agent modeling, Personality traits, Conversational agents.

Abstract: We present in this article an architecture implementing personality traits from the FFM/NEO PI-R taxonomy as influence operators upon the rational decision making process of dialogical agents. The objective is to separate designer-dependent resources (traits taxonomies, influence operators, behaviors/operators links) from the core part of the computational implementation (the personality engine). Through a case study, we show how our approach makes it easier to combine various resources and to observe various scenarios within a single framework.

1 INTRODUCTION

Designing virtual humans or agents to be used as long-term companions require them to display a believable behavior which remains consistent over time. In psychology, the concept of personality trait (Cattell et al., 1970; Goldberg, 1990) is defined as an habitual pattern of behavior or emotion, and therefore provides an appropriate theoretical foundation to build upon to reach the aforementioned goal. Once personality traits have been identified (or designed, in the case of an artificial agent), it is possible to anticipate (or define) their influence, in order to know extent how one will usually react in a particular situation: not only from an emotional perspective, with works from (Ortony et al., 1988) often used to implement psychological phenomena into artificial agents, but also from a rational point of view (Damasio, 1994), as studied for artificial agents by Rousseau and Hayes-Roth (Rousseau, 1996; Rousseau and Hayes-Roth, 1996).

However, most research works on the computational implementation of psychological phenomena (*cf.* discussion in Section 4) usually fail to take into account two key notions: *coverage*, as they often focus only on a small subset of psychological phenomena (*e.g.* considering few traits), and *comprehensiveness*, because they resort to procedural implementations (*e.g.* hard-coded rules) therefore excluding experts (*i.e.* psychologists) from the agent's behavior design process. The work presented in this paper aims at addressing those two restrictions.

Coverage issues: A key question regarding the principle of influence lies in the actual *extent* of the psychological influence over the reasoning. Typically, artificial agents focus on distinct subsets of domain-dependent psychological notions (*e.g.* a poker player (Findler, 1977)). However, the growing interest in conversational agents (Cassell et al., 2000) opens new perspectives where psychological notions become first class citizens (*e.g.* a different approach to poker player (Koda and Maes, 1996)), thus leading to a need not only for larger psychological domains, but also for a more generic way to handle them.

Research works in psychology offer several personality traits taxonomies, but because such taxonomies try to cover a large set of aspects of the personality of a person, they are in turn too general from a computational viewpoint: Cattell's 16 personality factors (Cattell et al., 1970), only 5 large classes in a single level for the Five Factor Model (FFM) (Goldberg, 1992), and 30 bipolar classes in the two-level Revised NEO Personality Inventory (FFM/NEO PI-R), which extends FFM. It is therefore difficult to define a precise interpretation of their classes in terms of operators over the rational process of agents, even for FFM/NEO PI-R, the most fine-grained of the commonly used taxonomies. For example, this led us (Sansonnet and Bouchet, 2010) to propose an extended version of FFM/NEO PI-R with a third level of so-called *behavioral schemes* that increases the precision in terms of classes (69 bipolar schemes) and lexical semantics (each scheme being defined by a set of actual behaviors).

Comprehensiveness issues: Assuming that a well-grounded and precise taxonomy of personality is available, a second question follows: what kind of influence operators over the agent’s process can be elicited from and associated with the taxonomy classes? Some works have proposed models describing how influences operators can be associated with taxonomy classes (*cf.* examples in Section 4), proving the feasibility of such an approach on case studies, but they are usually based on small subsets of arbitrarily chosen psychological behaviors. Therefore there is a need for a more comprehensive approach to the systematic implementation of *complete* personality traits domains (*e.g.* covering FFM) onto the rational process of artificial agents¹, with two main requirements:

- *Computational implementation:* no complete, orthogonal, and approved set of operators that would apply to main agent frameworks (from different fields such as artificial intelligence, multi-agents systems or intelligent virtual agents) currently exists. A *modular* and flexible approach is needed, to allow subsets of operators to be implemented in distinct frameworks.
- *Psychological relevance:* we need a model of relationships between classes and operators approved by psychologists. It would require a *declarative* approach, where distinct models of relationships could be shared by psychologists for experimentation and discussion, thus excluding procedural encoding.

Managing influences with personality engines: We propose an approach in which resources are both application-dependent and designer-dependent representations, and where the *personality engines* combine those resources to implement actual scenarios. This concept of personality engine allows to easily implement and test various psychological hypotheses through resource combination, but also to apply them to a wide variety of application domains for experimentation and evaluation purposes.

This article is organized as follows: in section 2, we introduce the enriched taxonomy of personality we have chosen to use in this study, show how it can be used on an example from the literature and introduce the concepts necessary to define a personality engine. Section 3 presents a case study using a simplified world of dialogical agents, shows how those agents can be provided with a personality engine affecting the way they communicate with each other, and demonstrates how it can be used to implement the example from section 2. Section 4 compares our approach to other attempts at implementing personality features in agents.

¹Complete coverage has been attempted for emotions, as in OCC (Ortony et al., 1988).

Table 1: Two-level FFM/NEO PI-R taxonomy

FFM Traits	FFM/NEO PI-R facets (each symbol includes a + and a - (antonym) pole)
Openness	Fantasy, Aesthetics, Feelings, Actions, Ideas, Values
Conscientiousness	Competence, Orderliness, Dutifulness, Achievement-striving, Self-discipline, Deliberation
Extraversion	Warmth, Gregariousness, Assertiveness, Activity, Excitement-seeking, Positive-emotions
Agreeableness	Trust, Straightforwardness, Altruism, Compliance, Modesty, Tender-mindedness
Neuroticism	Anxiety, Angry-Hostility, Depression, Self-consciousness, Impulsiveness, Vulnerability

2 THE PERSONALITY ENGINE

2.1 An enriched personality domain

2.1.1 The traditional FFM/NEO PI-R taxonomy

Several theoretical approach to study human personality have been developed over years: Freudian psychoanalysis, types and traits, Maslow and Rogers’ humanistic psychology, Bandura’s social-cognitive theory, etc. Among them, personality traits have been widely used as a ground for studies in affective computing (Rizzo et al., 1997) and cognitive agents (Gratch and Marsella, 2004). We will therefore rely on them and focus on the FFM/NEO PI-R taxonomy (Goldberg, 1981), which is the most prominent one in the context of computational studies (*cf.* (John et al., 2008)). The FFM/NEO PI-R taxonomy is made of five classes of psychological behaviors, also called O.C.E.A.N. traits. Each FFM trait is divided into six sub-classes (called *facets*) resulting in 30 bipolar² positions (Costa and McCrae, 1992), listed in Table 1. The semantics of each facet is intuitively defined by a unique gloss³, *e.g.* facet *Fantasy* is defined by “receptivity to the inner world of imagination” and *Aesthetics* by “appreciation of art and beauty”.

2.1.2 The enriched FFM/NEO PI-R/BS taxonomy

The FFM/NEO PI-R taxonomy has the advantage of being grounded on state of the art research in psychology, which allows us to safely consider that it covers a large part of the domain of a person’s personality

²Each facet has a positive (*resp.* negative) pole noted + (*resp.* −) associated with the concept (*resp.* the antonym of the concept). Facets are usually referred to using the name of their + pole.

³A gloss is a short natural language phrase defining intuitively a lexical semantics sense, as found in dictionaries or in WordNet synsets (Fellbaum, 1998).

traits. However, when one is interested in the computational expression of psychological phenomena such as personality traits, the facet definitions (based on a unique gloss per facet as in the aforementioned examples) are too general from two complementary points of view:

1) They can cover a large set of psychological behaviors so that scripting the psychology of a character can be imprecise. A third level, breaking down facets into smaller subsets would facilitate an association with more specific behaviors.

2) Definitions are so general that defining a precise functional relation between facets and influence operators can be difficult, which also encourages to go towards breaking down facets into more specific psychological behaviors.

These considerations led us to rely on an enriched three-level taxonomy of FFM/NEO PI-R, called FFM/NEO PI-R/BS (Bouchet and Sansonnet, 2010) and available on the Web⁴, in which each facet of FFM/NEO PI-R, is decomposed in so-called behavioral schemes (or *schemes* in short). It extends FFM/NEO PI-R by associating glosses to the senses of a large set of 1 055 personality adjectives, using the WordNet database (Fellbaum, 1998), completed and aligned with 300 Goldberg's questionnaire so-called *q-items*⁵, and for each FFM/NEO PI-R position, glosses and items have been clustered into sets⁶ of congruent operational behaviors: the schemes.

Quantitatively, FFM/NEO PI-R/BS taxonomy features: $N_{\text{facet}} = 30$, $N_{\text{gloss}} = 766$, $N_{\text{scheme}} = 69$, $\bar{N}_{\text{glosses/facet}} = 26$ and $\bar{N}_{\text{schemes/facet}} = 2.3$.

2.1.3 Example: defining a personality profile into the FFM/NEO PI-R/BS taxonomy

We propose to consider an example taken from CyberCafe in Rousseau and Hayes-Roth (1996), in which several characters who endorse the same interactional role of a *waiter* (w_i) have distinct psychological profiles $P(w_i)$, entailing distinct psychological behaviors $B(w_i)$ such as:

$P(w_1)$ realistic, insecure, introverted, passive, secretive
 $B(w_1)$ Such a waiter does and says as little as he can
 $P(w_2)$ imaginative, dominant, extroverted, active, open
 $B(w_2)$ This waiter takes initiative, comes to the customer without being asked for, talks much

Considering the psychological profile $P(w_1)$ of waiter w_1 , it can easily be transposed onto the FFM/NEO PI-

R/BS taxonomy in terms of scheme activations (formal definition is given in Section 2.2.3):

$$P'(w_1) = \{ \begin{array}{ll} \text{realistic} & \Rightarrow \text{O-fantasy-PRACTICAL;} \\ \text{insecure} & \Rightarrow \text{C-competence-INSECURE;} \\ \text{introverted} & \Rightarrow \text{E*}(-\text{COLD, -NONGOSSIPMONGER, -SOLITARY,} \\ & \text{-UNCOMMUNICATIVE, -UNCHARISMATIC,} \\ & \text{-DISCRET, -SUBMISSIVE, -PLEADING, -} \\ & \text{LANGUID, -APATHETIC, -ASCETIC, -BLASE);} \\ \text{passive} & \Rightarrow \text{E-activity-APATHETIC;} \\ \text{secretive} & \Rightarrow \text{A-trust-SECRETIVE} \end{array} \}$$

where elements of $P(w_1)$ are transposed in order, separated by ',' in $P'(w_1)$. We can notice that this profile mainly activates negative poles and that a FFM/NEO PI-R/BS scheme can easily be found to correspond to each P trait (which means that P traits are more schemes than actual FFM traits or FFM/NEO PI-R facets). The only exception is *introverted*, which is associated to the whole FFM trait *-Extraversion*, thus entailing 12 schemes, which adds precision. The same remarks apply to $P(w_2)$ but for the activation of positive poles; it is actually likely that $P(w_1)$ and $P(w_2)$ were hand-built.

$P'(w_1)$ offers a more systematic positioning in FFM/NEO PI-R and a more precise behavioral definition because the definition $B(w_1)$ is replaced with the glosses associated with the activated schemes in FFM/NEO PI-R/BS. For example, *-PRACTICAL* is defined by the WordNet glosses (N_i) and Goldberg's q-items (Q_i) associated to it:

N618	guided by practical experience and observation rather than theory
N626	aware or expressing awareness of things as they really are
N788	freed from illusion
N1232	concerned with the world or worldly matters
N795	sensible and practical
Q6	Spend time reflecting on things
Q7	Seldom daydream
Q8	Do not have a good imagination
Q9	Seldom get lost in thought

In summary, FFM/NEO PI-R/BS offers a precise grounding for personality description: not only it covers the eight classes proposed in Cyber-cafe (Rousseau, 1996), but it also enables a more precise and practical behavioral description, which justifies our decision to use it in the following sections.

2.2 Architecture of a personality engine

2.2.1 Personality engine structure

We define a personality engine PE as a 5-tuple such as $PE = \langle O, W, T, \Omega_W, M \rangle$ where:

– O is a *personality ontology* that enables precise descriptions of personalities. We will use in this pa-

⁴<http://perso.limsi.fr/jps/research/rnb/toolkit/taxo-glosses/taxo.htm>

⁵<http://pip.ori.org/newNEOKey.htm>

⁶Like facets, schemes are bipolar and are often referred to by their +pole.

per the set Σ of bipolar schemes from FFM/NEO PI-R/BS (described in Section 2.1.2), thus $|\Sigma| = 69$. The subset of positive (*resp.* negative) positions is denoted $+\Sigma$ (*resp.* $-\Sigma$), and their union is $\pm\Sigma$ such as $\pm\Sigma = +\Sigma \cup -\Sigma$ and $|\pm\Sigma| = 138$;

– $\mathbb{W}$ is an agent *world* model that includes: their internal structure $\mathbb{W}_s$; their external communication protocols $\mathbb{W}_c$; their rational decision making process $\mathbb{W}_r$. For example, a BDI-based model or a more specific one, such as the one defined in Section 3.1.1;

– $\mathbb{T}$ is an *application topic* enabling the instantiation of $\mathbb{W}$ in a particular case;

– $\Omega_{\mathbb{W}}$ is a set of *influence operators* over $\mathbb{W}_r \cup \mathbb{W}_c = \mathbb{W}_{rc}$;

– $\mathbb{M}$ is an *activation matrix*, establishing a relation over $\pm\Sigma \times \Omega_{\mathbb{W}}$.

$\mathbb{O}$, $\mathbb{W}$ and $\mathbb{T}$ are considered as given resources, whereas $\Omega_{\mathbb{W}}$ and $\mathbb{M}$ must be elicited from the resources, as explained in 2.2.2 and 2.2.3.

2.2.2 Influence operators elicitation

Given an agent model $\mathbb{W}$, influence operators are meta rules $\omega \in \Omega_{\mathbb{W}}$ controlling or altering the non structural parts of $\mathbb{W}$, *i.e.* $\mathbb{W}_{rc}$.

Example: Let us consider some plan in $\mathbb{W}_r$ containing the expression $e = \text{PAR}[a_1, a_2, a_3]$, which is a set of three actions to be executed in no particular order (like operator PAR of CSP). One can define the rule $\omega_1 = \text{PAR} \rightarrow \text{SEQ}$ which, applied to e , can intuitively stand for an indication to an agent to execute its actions routinely (and correctly). On the contrary, a rule $\omega_2 = \text{SEQ} \rightarrow \text{PAR}$ could stand for a disorderly agent (and sometimes lead to incorrect executions of the plan).

This simple example shows that whenever, formally, any rule over $\mathbb{W}_{rc}$ is an influence operator, only those that could be interpreted in terms of psychological behaviors are actually relevant. Consequently, one has to consider operator elicitation as an operation from $\mathbb{W}_{rc} \times \mathbb{O} \mapsto \Omega_{\mathbb{W}}$ rather than $\mathbb{W}_{rc} \mapsto \Omega_{\mathbb{W}}$.

The definition of an algorithm that takes a couple of resources $\mathbb{W}$ and $\mathbb{O}$ and automatically produces the⁷ set $\Omega_{\mathbb{W}}$ is still an open question. For the time being, we have to restrict to hand-built operators sets, which are de facto designer(s)-dependent. The notion of personality engine makes it possible to handle the management of this diversity (*e.g.* distinct propositions PE_i , based over the same $\mathbb{W}$ and/or $\mathbb{O}$, can be tested and systematically compared). An example of operator elicitation is detailed in Section 3.2.1.

⁷using ‘the’ raises issues of existence (no possible influences found) and unicity (several distinct sets found thus prompting an order relation).

Operators intensity and direction: Operators like PAR and SEQ, are activated straightforwardly: they are applied or not. However, various operators can be activated in more complex manners through argument passing. We will consider two frequent cases:

– An intensity is given, *cf.* activation levels in Table 3;

– Operators also working in reverse or antonym mode can be given a direction (*e.g.* operator ω_{-safe} in Section 2.3.1).

2.2.3 Activation matrix elicitation

Once given the set schemes $\sigma \in \pm\Sigma$ and a set of influence operators $\omega \in \Omega_{\mathbb{W}}$, the designer(s) of a particular processing engine must elicit how $\pm\sigma_i$ are linked to ω_i , that is which schemes activate which operators. This relation, which is again designer-dependent, is established by a multi-valued matrix $\mathbb{M}$ of so-called *activation levels* $\lambda_{i,j}$ such that $\mathbb{M} = \pm\Sigma \times \Omega_{\mathbb{W}}$. Elements $\lambda_{i,j}$ of $\mathbb{M}$ have the following values and conventions:

- 2 activate operator with strong force
- 1 activate operator with moderate force
- 0 the operator is deactivated
- 1 activate antonym operator (if it exists) with moderate force
- 2 activate antonym operator (if it exists) with strong force

2.3 Instantiating personality engines

Once given a particular personality engine PE_0 , one has a symbolic structure that can be instantiated into actual situations varying from two main points of view: application topics and personality profiles.

2.3.1 Application topics

Let $\mathbb{T}_0$ be a particular topic providing a set of available actions $\alpha_i \in \mathbb{A}(\mathbb{T}_0)$. The topic also provides influence operators of PE_0 with application-dependent information about α_i . For example, let ω_{+safe} be an operator that sorts a set of actions from the safest to the least safe : $\omega_{+safe} \doteq \text{Sort}(\{\alpha_i\}, \prec_{danger})$. To be operational, operator ω_{+safe} requires topic $\mathbb{T}_0$ to provide a measure function $\mu_{danger} : \mathbb{A}(\mathbb{T}_0) \mapsto [0, 1]$. Operator ω_{+safe} has an antonym, ω_{-safe} , that sorts actions in reverse order.

2.3.2 Personality profiles

Intuitively, personality profiles are often defined as sets of adjectives/adverbs describing the behavior of a person. For instance, in the Cybercafe example (*cf.*

Section 2.1.3), personality profile $P(w_1)$ was first defined with a set of common *words*: {realistic, insecure, introverted, passive, secretive}. The research about personality trait taxonomies enables more precise definitions that use a mapping in terms of sets of well-grounded *concepts*, like $P'(w_1)$. Using FFM/NEO PI-R/BS prompts the following definition:

Given an individual x , its personality profile $P(x)$ can be defined as a set of $|\Sigma|$ functions $p(\sigma_i) : \Sigma \rightarrow \{+, \asymp, -\}$ where:

- $\asymp$ means that with regard to scheme σ_i , person x 's behavior is not significantly deviant from an average behavior;
- $+$ means x 's behavior is deviant from average according to $+$ -pole;
- $-$ means x 's behavior is deviant from average according to $-$ -pole.

Notation: When one considers the 69 schemes of Σ , people tend to exhibit an average behavior for most of them. Consequently $P(x)$ is often a scarce vector with most elements valued with $\asymp$, so $P(x)$ is preferably given as a set of non $\asymp$ schemes. For example, Paul's personality will be denoted in short: $P(\text{Paul}) = \{-\text{HARDWORKER}, -\text{ATTENTIVE}, \text{HARMLESS}, \text{EMPATHIC}, -\text{SHOWY}\}$, ignoring the 64 other schemes for which his behavior doesn't stand out.

3 CASE STUDY

In this section we present a case study showing how personality engines can be defined, then instantiated in actual situations. To support the eliciting process of influence operators, one must choose an application model, for example, to focus on agents' communication, well-used KQML, ACL-FIPA or BDI with logics (KGP, 2APL, Golog-based etc.).

3.1 TALKINGS: a typical world of dialogical agents

We consider here a simplified model, called TALKINGS (a simple world of agents interacting through message passing), that allows a comprehensive presentation of our approach. For this example, we have chosen to focus on conversational agents, a fast growing application domain. Consequently, we will consider traits and operators associated with social and dialogical aspects of the agents, which cover about 55% of the FFM/NEO PI-R/BS schemes (Sansonnet and Bouchet, 2010). The process described here is complementary to non-dialogical aspects of the agents studied in (Bouchet and Sansonnet, 2011)).

3.1.1 Agent model

Let TALKINGS be an actual world composed of physical or abstract entities, which is accessed through a representation of its entities into a symbolic model $\mathcal{M}$. An entity $e_i \in \mathcal{M}$ is defined in $\mathcal{L}_M$, its associated language of description, as a set of rule-based definitions of the general form $D_i = \text{left part} \mapsto \text{right part}$ such that $\forall e_i \in \mathcal{L}_M; e_i = \{D_i\}$.

Agents $a_i \in \mathcal{A}$ represent dialogical entities of $\mathcal{M}$ that can perform practical reasoning. An agent $a_i \in \mathcal{A}$ is defined as a 5-tuple $\langle id, K, S, \Phi, \Psi \rangle$ where:

- id is a string providing a unique identifier for the agent;
- Knowledge base $K = k_i \in \mathcal{L}_k$ is a set of propositions over $\mathcal{M}$;
- Social base S is the set of roles endorsed by the agent (over TALKINGS, or relatively to another agent of TALKINGS);
- Feature base Φ is the set of physical attributes of the agent (to simplify, Φ will not be considered further);
- Psychological base $\Psi = \Psi_T \cup \Psi_M$ is a set of static traits Ψ_T and dynamic moods Ψ_M (dynamic moods are out of scope here since we focus on personality associated with static traits).

3.1.2 Message structure

Collectives c_i of TALKINGS agents support the operation $\text{SEND}[t, a, \{b_i\}, m]$ enabling the transfer of message m at turn t between the sender agent a and one or more receiver agents $\{b_i\}$. In the following, we restrict this definition to interactions between the couple of agents $a \Leftrightarrow b$ (in the following, a denotes the so-called *speaker* and b its *interlocutor*) hence considering operations of the form $\text{SEND}[t, a, b, m]$. A message m into such SEND operations contains four expressions, explicitly stated by speaker agent a towards interlocutor agent b :

$$m = \langle \text{Reaction}, \text{Proaction}, \text{Forces}, \text{Content} \rangle$$

Reaction is the attitude that a adopts, and expresses explicitly, in reaction to its own evaluation of the previous message from b at turn $t - 1$. Reactions are organized on a $-/+$ scale, ranging from total disagreement (noted *No*) to total agreement (noted *Yes*). The first message of the first turn of a session has an empty reaction (noted $-$).

Proaction is the main attitude stated by a towards b . Two main proactive attitudes are considered, according to the direction of the intention of a :

- *Ask*, represented as $a \xleftarrow{A \text{ Content}} b$, where agent a sends a query to b about Content;

Table 2: Intuitive semantics of the levels of activation of the message operators.

levels		Operators			Activation levels λ						
1	2	CODE	Label	Definition	-2	-1	0	1	2	Range	
Proaction	Explicit	A	Ask	probability for the agent to tend to use <i>Ask</i> or <i>Propose</i>	-	-	none	ask if needed	ask even if not needed	[[0, 2]]	
		P	Propose		-	-	none	propose if needed	propose even if not needed		
		D	Dominance	probability for the agent to use the force or its antonym	inferior	supporter	none	equal	superior	[[−2, 2]]	
		F	Feeling		aggressive	cold	none	polite	warm		
		M	Motivation		show false motive	hide motive	none	motive if needed	motive even if not needed		
		I	Incentive		menace		none	promise			
	Implicit	G	Guess	capacity of the agent to perceive other agents in terms of their rational processes, their mental states, etc.	perceive false	do not perceive	none	perceive if explicit	perceive even if not explicit		
		C	Conflict	attitude of the agent about risking to provoke conflicts	like conflicts	accept conflicts	none	dislike conflicts	avoid any conflict		
		S	Sincerity	sincerity of the agent about all parts of a sent message	tell false facts	actively hide facts	none	frank	very/too frank		
Reaction	Explicit	A+	Reaction	typical reaction to an <i>Ask</i> or <i>Propose</i> depending on the global evaluation by <i>b</i> of the Forces expressed by <i>a</i>	always no	yes but with protest	none	yes but can be conditional	always yes	[[−2, 2]]	
		A-									
		P+									
		P-									
	Implicit	B	Bond	reaction to results of <i>G</i> (<i>e.g.</i> perceiving <i>a</i> is sad, <i>b</i> will: +) feel sad; 0) not care; -) feel happy)	bond to reverse	do not bond	none	bond if needed	bond even if not needed		
		N	Negotiate	reaction in the management of conflicts existing and explicit	increase	sustain	none	settle	always yield		

Note: level 2 (resp. -2) includes level 1 (resp. -1), i.e. it can exhibit behaviors of level 1 (resp. -1)

– *Propose*, represented as $a \xrightarrow{P \text{ Content}} b$, where agent *a* sends a proposition to *b* about Content.

Forces are optional modalities of proaction operators (A|P), explicitly expressed by *a*, in order to contribute to the expected success of the message. A message from *a* is considered successful when in reply from *b*, the reaction of *b* is positive and the proaction of *b* is relevant to *a*. We consider four distinct forces, each organized as a bipolar -/+ scale:

- *Dominance* ranges from force -submissive to +dominant, which can modalize operators A|P, e.g. A -submissive can be viewed as begging and A +dominant as requesting.
- *Feelings* ranges from force -aggressive to +affective, which can accompany operators A|P.
- *Motivation* ranges from force -hide to +open. An agent using open force explains clearly and frankly the rational motive(s) of the sending. Conversely, the agent can try to hide its rational motives or even to express untrue motives.
- *Incentive* ranges from force -menace to +promising. An agent *a* using +promise force attempts at facilitating the success of its message by providing rational positive reasons for *b* to react positively to it, or by addressing direct rewards. Conversely, *a* can try to obtain agreement from *b* through -menace (e.g. by stating

rational negative outcomes for *b* if it disagrees) or by addressing direct threats.

Content is the body of the message, that is the object of the proaction. Five main classes of objects are considered:

- *Knowledge* is a fact $k_i \in \mathcal{L}_k$;
- *Action* is an operation upon the world. For example, $A a(x)$ means *a* asks *b* to execute $a(x)$, while $P a(x)$ means *a* intends to execute $a(x)$;
- *Resource* is an entity in the world that can be possessed and transferred;
- *Norm* describes rights or duties of agents in a given collective c_i ;
- *Emotion* describes a personal mental state (e.g. mood) or an interpersonal affective relationship.

With these definitions, the structure of a message *m* can be represented as:

– $|Yes|No \times A|P \times [D][F][M][I] \times k|a|r|n|e$
 where | separates alternatives, [] embraces optional forces, k, a, r, n, e are the five types of content and $\times$ is the Cartesian product, thus defining the message domain. A turn *t* is a couple $\langle \text{SEND}[t, a, b, m], \text{SEND}[t, b, a, m'] \rangle$ where m' is the reply to *m*. A simple interactional session is a sequence of turns; more complex sessions can include sub-

sessions (called threads) *e.g.* in case of conditional reactions.

3.2 Building a personality engine in TALKINGS

3.2.1 Eliciting influence operators

Considering the previous agent's model (*i.e.* $\mathbb{W} = \text{TALKINGS}$), it is possible to associate with the model a set of influence operators Ω_{TALKINGS} that define meta control over the rational decision making process of the agents $\mathbb{W}_r$ and over the message passing process $\mathbb{W}_c$. We will focus here on the operation of building and sending messages, *i.e.* on $\mathbb{W}_c$. Browsing the model, described in section 3.1.1, we can define 15 operators organized in a 2×2 ontology, mirroring the model structure: at the first level of the ontology, influence operators on message passing can be divided into two main classes, *proaction* and *reaction*, and at the second level, we can distinguish for each class *implicit* and *explicit* operators. We therefore distinguish:

- *Explicit proaction* operators, which are expressed into messages.
- *Implicit proaction* operators, which are not explicitly expressed in messages but can influence the way messages are built and are related to the social capacities of the agent.
- *Explicit reaction* operators, which are expressed into messages, in terms of Yes/No reactions.
- *Implicit reaction* operators, which mirrors implicit proaction.

Table 2 gives a list of exhibited message operators together with an abridgment of their semantics associated with their activation levels λ , ranging on scales with discrete positions defined in Section 2.2.3.

While we have used a simplified communicating agent model, together with the description of the FFM/NEO PI-R/BS schemes, it was possible to exhibit 12 operators, defined and organized as in Table 2. In comparison, the eight “types of behaviors” similar to our operators (Perceiving, Reasoning, Learning, Deciding, Acting, Interacting, Revealing, Feeling) given in Cybercafe (Rousseau and Hayes-Roth, 1996) remain rather general, although some can be directly mapped onto TALKINGS operators such as Perceiving and Guess, Learning and Ask, Revealing and Motivation. Feeling would not be handled here since we consider interactions only and not internal emotions.

3.2.2 Establishing an activation matrix

Given the set $\pm\Sigma$ and the set Ω_{TALKINGS} of elicited operators in the case study TALKINGS, it is possible

to define an activation matrix M_{TALKINGS} , which establishes the relationships between the schemes and the operators. Table 3 shows an excerpt of a proposition for M_{TALKINGS} (from the 138 schemes of $\pm\Sigma$, we display only the 10 schemes used in the example of Section 3.2.3). Not to overload Table 3, activation values $\lambda_{i,j}$ that are associated with an average behavior are factorized in headline “Generic agent” and represented as empty cells.

3.2.3 Example of personality scripting

As an example of instantiation of the personality engine defined for TALKINGS, we consider $P'(w_1)$ from the Cybercafe example (*cf.* Section 2.1.3). For simplification purposes, the 12 schemes associated with adjective ‘introverted’ are coerced into a single one $E_{\text{warmth-COLD}}$ (first arbitrarily chosen) thus prompting a new profile: $P''(w_1) = \{ O_{\text{fantasy-PRACTICAL}}; C_{\text{competence-INSECURE}}; E_{\text{warmth-COLD}}; E_{\text{activity-APATHETIC}}; A_{\text{trust-SECRETIVE}} \}$. Respectively, for waiter w_2 we have: $P''(w_2) = \{ O_{\text{fantasy+CREATIVE}}; E_{\text{assertiveness+DOMINEERING}}; E_{\text{warmth+FRIENDLY}}^8; E_{\text{activity+ACTIVE}}; E_{\text{fantasy+IDEALISTIC}}^9 \}$.

Values of activation levels associated with $P''(w_1)$ and $P''(w_2)$ in M_{TALKINGS} are given in Table 3. For example, crossing Tables 2 and 3, it is possible to identify the influences of one of the schemes of $P''(w_1)$ *e.g.* $A_{\text{trust-SECRETIVE}}$ (last line of Table 3):

A/0	don't ask explicitly (while average behavior would be 1: ask if needed by the rational process)
P/0	don't propose explicitly (idem)
F/0	no sensibility to inner feelings activated (idem)
M/-1	hide one's own motives
I/0	usage of positive or negative incentives over others deactivated
A/-2	react explicitly always by a rejection when asked with a force considered negative
A+/-1	react explicitly positively but with protest, when asked with a force considered positive
B/0	no bond positive or negative is activated (while the average behavior would be 1: bond if needed by the rational process <i>e.g.</i> in social condolences)

Operators A, P, F, I, B are controlled via deactivation ($\lambda = 0$). Actually average behavior often uses ($\lambda = 1$)

One can make the following remarks:

R1 Over the set of 12 operators in Table 3, scheme +PRACTICAL is not distinct from Generic agent. This is consistent with the fact that Bratman's agents implement an implicit personality close to scheme +PRACTICAL and be viewed as a particular case.

⁸First scheme chosen with same rule in $P''(w_1)$.

⁹Again, first scheme of trait Open chosen.

Table 3: Excerpt from Activation matrix $M_{TALKINGS}$. When $\lambda_{i,j} = \emptyset$ then $\lambda_{i,j} = \text{GenericAgent}_j$

			Proaction									Reaction								
Operator code			A	P	D	F	M	I	G	C	S	A+	A-	P+	P-	B	N			
Value range			02	02	2-2	2-2	2-2	2-2	2-2	2-2	2-2	2-2						2-2		
Generic agent			1	1	1	1	1	1	1	1	1	1	-1	1	-1	1	1	Waiters		
T	Facet	Scheme																w ₁	w ₂	
O	fantasy	-PRACTICAL																	*	
O	fantasy	+IDEALISTIC	2		2		2	2	-1		2	2		2		0			*	
O	fantasy	+CREATIVE	2		2				-2											*
C	competence	-INSECURE	2	0	-2		2		0	2		1		1		2		*		
E	warmth	+FRIENDLY			-1		2		2		2		2		2		2		*	
E	warmth	-COLD			-1				-1		-1	-1	-1						*	
E	assertiv.	+DOMINEER.	2	2		-1		0	-1	-1		-1	-2		-2		-1		*	
E	activity	+ACTIVE	2						-1		2		2				-1		*	
E	activity	-APATHETIC	0	0	-2		0		0	0	1		1		-1		2	*		
A	trust	-SECRETIVE			0		-1		0			-1		-2		0		*		

-PRACTICAL is the antonym pole of scheme +IDEALISTIC *resp.* +FRIENDLY/-COLD, +ACTIVE/-APATHETIC.

R2 All lines of Table 3 are distinct, entailing that all schemes are distinct concepts with distinct sets of influences.

R3 It happens that profiles of the Cybercafe waiters, $P''(w_1)$ and $P''(w_2)$ activate exclusive schemes (*). Their definition is not always consistent, meaning that some schemes are activated by contradictory levels *e.g.* 1 and -2. In theory, when a personality is scripted, nothing prevents from defining conflicting activations of the same operator: our approach makes it easier to automatically check for such cases and to handle them manually or automatically, according to an order relation possibly provided by psychologists.

3.3 Discussion

Relevance and completeness of the operators The process of operator elicitation ensures that all operators defined in trait Conscientiousness are relevant. For example, in the case study above, because they are synthesized from scheme glosses, they are activated in a non trivial manner at least once¹⁰ (*i.e.* $\forall \sigma \in M_{TALKINGS}, \exists j$ such that $\lambda_{i,j} \neq \text{generic-agent}(i)$).

Conversely, the elicitation process does not ensure that all possible operators are found; from a psychological point of view this is not yet attainable. Actually this issue is in support of our approach that is based on the state of the art of the coverage of the domain of the psychology of a person, that is to say trait taxonomies, in particular FFM/NEO PI-R. Moreover, the refined version FFM/NEO PI-R/BS, grounded

¹⁰Except for first line of table 3 (O fantasy -practical), which is similar to a line generic-agent as this trait can be viewed as Bratman’s notion of practical reason (1987).

on large ascertained lexical resources (*e.g.* WordNet), covers according to the state of current literature, the effective behaviors that are associated with personality traits, hence restricting the risk of silence.

Validation of activation matrix values Weights $\lambda_{i,j} \in M_{TALKINGS}$ are set by annotators. This results in 1) inter annotator quantitative differences that can be partly controlled with statistic tools acting over the annotating group; 2) qualitative controversies between computer science experts and psychologists. The proposed approach has the virtue of putting into light the essential issue of those qualitative controversies, usually embedded in the programming process of the procedural approaches, listed in Section 4. In our case, the use of a declarative method, through a matrix of activation levels instead of procedural rules, increases the comprehensiveness and the tracking of the traits/behaviors association. Moreover, the declarative approach clarifies the discussions with psychologists, who *in fine* must validate the decisions.

Evaluation of the model In this paper we propose an approach for handling the phenomena, stated in the literature, of personality traits influence over plans and actions. Our purpose is not the direct evaluation of a particular model (composed of: a specific rational model, a specific set of influence operators and a specific set of activation levels) through an experimentation. Here we pursue a double objective:

1. present a proof of concept of the principle of influence: ‘points of influence’ actually exist in the rational decision making process;
2. propose a method that is a) generic *i.e.* not designed for a small set of specific traits but covering

a large domain of the personality of a person; and b) declarative *i.e.* using explicit levels instead of embedded rules.

For example, a consequence is that Table 2 and Table 3 must be viewed as instances of our approach. As such, they need to be evaluated through proper experiments, but which are beyond the scope of this paper.

4 RELATED WORKS

Since works of Rousseau and Hayes-Roth (1996), extensive research has been undertaken, especially recently, involving both psychological phenomena and artificial agents in at least four communities: rational agents, multi-agents systems, conversational agents and affective computing.

Gratch and Marsella (2004) have implemented a psychological model, mainly dedicated to emotions, based on traditional SOAR architecture, but most authors have proposed improvements of BDI architectures exhibiting both rational reasoning modules and psychological reasoning modules (Lim et al., 2008). For example, the eBDI model (Jiang et al., 2007) implements emotions in a BDI framework, in which they give a good introduction about the necessity to implement emotions into rational agents. Indeed, BDI architectures offer an open and flexible engine (the deliberation cycle), for example using tools like 2APL (Dastani, 2008), which is why we rely on it for the support of the framework that underlies this study.

However our approach is distinct from most studies using BDI engines, mainly because in those studies the psychology of the agent is based on *dynamic* mental states (like moods and affects, as in Section 3.1.1), which influence the bodily (facial and gestural) expression of *emotions*, but they have no or little impact upon the decision making process of the agent, especially for controlling conversational strategies. Instead, in our approach the *static* features of the personality of an agent are expressed through its influences upon operational behaviors.

Using the BDI platform JACK (Howden et al., 2001), CoJACK (Evertsz et al., 2008) provides an additional layers which intends to simulate physiological human constraints like the duration taken for cognition, working memory limitations (*e.g.* “loosing a belief” if the activation is low or “forgetting the next step” of a procedure), fuzzy retrieval of beliefs, limited attention or the use of moderators to alter cognition. A similar approach is taken for conversational agents in PMFserv (Silverman et al., 2006).

However, in these studies, authors focus on the influence of *physical* or *cognitive* capacities over the de-

liberation cycle but not on actual psychological phenomena like moods or traits.

Closer to our work, Malatesta et al. (2007) use traits to create different expressions of emotions, especially by influencing the appraisal part of the OCC theory (Ortony et al., 1988). They focus on how agents evaluate the results of their actions and of external events, whereas we focus on the way they perform a task. In the same way, Rizzo et al. (1997) have shown that goals and plans can be used to represent a character’s personality in an efficient way, by attributing specific behaviors to the pursuit of each goal. Personality traits are used to choose between the multiple goals of a BDI agent (*i.e.* traits influence Desires). Once chosen, goals are planned and executed directly.

However, in our case, traits operate on already planned goals (*i.e.* traits influence Intentions). This remark also applies to (McRorie et al., 2009), based on the architecture of conversational agent GRETA (Pelachaud, 2000), which involves models of personality for the expression of emotions (face, gesture, etc.) and to the FATIMA architecture (Doce et al., 2010) stemming from (Pelachaud, 2000), which implements personality traits.

Finally, all these studies share the same approach to psychology, each of them focusing on particular capacities or particular traits. They do not attempt to cover a whole domain, hence they are not concerned with managing and comprehensiveness issues.

5 CONCLUSION

We have shown in this article an approach based on personality engines which provides three main advantages: firstly, it reduces and reifies author/designer-dependent parts in only three main kinds of resources: trait ontologies, sets of influence operators and activation matrices. Secondly, it defines a process for designing the resources and for implementing, in a declarative way (activation matrix), personality influences in dialogical agents. Finally, it proposes an architecture where these resources can be flexibly combined (*cf.* section 3.2) and easily observed (*cf.* section 3.2.3).

We intend to extend this work in two main directions: by eliciting operators over outstanding BDI agent frameworks to demonstrate its independence over the framework used, and by experimenting with scenarios supervised by psychologists. The perception of the implemented agent’s personality by human users could be evaluated post-interaction with questionnaires such as the Agent Persona Instrument (Baylor and Ryu, 2003).

REFERENCES

- Baylor, A. and Ryu, J. (2003). The api (agent persona instrument) for assessing pedagogical agent persona. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, volume 2003, pages 448–451.
- Bouchet, F. and Sansonnet, J. P. (2010). Classification of wordnet personality adjectives in the NEO PI-R taxonomy. In *Fourth Workshop on Animated Conversational Agents, WACA 2010*, pages 83–90, Lille, France.
- Bouchet, F. and Sansonnet, J. P. (2011). Influence of personality traits on the rational process of cognitive agents. In *The 2011 IEEE/WIC /ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Lyon, France.
- Bratman, M. E. (1987). *Intentions, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA.
- Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors (2000). *Embodied Conversational Agents*. MIT Press.
- Cattell, R. B., Eber, H. W., and Tatsuoka, M. M. (1970). *Handbook for the sixteen personality factor questionnaire (16 PF)*. Champaign, Illinois.
- Costa, P. T. and McCrae, R. R. (1992). *The NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.
- Damasio, A. R. (1994). *Descartes error: Emotion, reason and the human brain*. New York: G.P. Putnam's Sons.
- Dastani, M. (2008). 2APL: a practical agent programming language. In *AAMAS '08: The seventh international joint conference on Autonomous agents and multiagent systems*, volume 16, pages 214–248, Estoril, Portugal. Springer-Verlag.
- Doce, T., Dias, J., Prada, R., and Paiva, A. (2010). Creating individual agents through personality traits. In *Intelligent Virtual Agents (IVA 2010)*, volume 6356 of *LNAI*, pages 257–264, Philadelphia, PA. Springer-Verlag.
- Evertsz, R., Ritter, F. E., Busetta, P., and Pedrotti, M. (2008). Realistic behaviour variation in a BDI-based cognitive architecture. In *Proc. of SimTecT'08*, Melbourne, Australia.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.
- Findler, N. V. (1977). Studies in machine cognition using the game of poker. *Commun. ACM*, 20(4):230–245.
- Goldberg, L. R. (1981). Language and individual differences: The search for universal in personality lexicons. *Review of personality and social psychology*, 2:141–165.
- Goldberg, L. R. (1990). An alternative description of personality: The big-five factor structure. *Journal of Personality and Social Psychology*, 59:1216–1229.
- Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, 4:26–42.
- Gratch, J. and Marsella, S. (2004). A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269–306.
- Howden, N., Rannquist, R., Hodgson, A., and Lucas, A. (2001). Intelligent agents - summary of an agent infrastructure. In *Proc. of the 5th International Conference on Autonomous Agents*, Montreal.
- Jiang, H., Vidal, J. M., and Huhns, M. N. (2007). eBDI: an architecture for emotional agents. In *AAMAS '07: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pages 1–3, New York, NY, USA. ACM.
- John, O. P., Robins, R. W., and Pervin, L. A., editors (2008). *Handbook of Personality: Theory and Research*. The Guilford Press, 3rd edition.
- Koda, T. and Maes, P. (1996). Agents with faces: the effect of personification. pages 189–194.
- Lim, M., Dias, J., Aylett, R., and Paiva, A. (2008). Improving adaptiveness in autonomous characters. In Prendinger, H., Lester, J., and Ishizuka, M., editors, *Intelligent Virtual Agents*, volume 5208 of *Lecture Notes in Computer Science*, pages 348–355. Springer.
- Malatesta, L., Caridakis, G., Raouzaoui, A., and Karpouzis, K. (2007). Agent personality traits in virtual environments based on appraisal theory predictions. In *Artificial and Ambient Intelligence, Language, Speech and Gesture for Expressive Characters, achieve AISB'07*, Newcastle, UK.
- McRorie, M., Sneddon, I., de Sevin, E., Bevacqua, E., and Pelachaud, C. (2009). A model of personality and emotional traits. In *Intelligent Virtual Agents (IVA 2009)*, volume 5773 of *LNAI*, pages 27–33, Amsterdam, NL. Springer-Verlag.
- Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge, UK, Cambridge university press edition.
- Pelachaud, C. (2000). Some considerations about embodied agents. In *Int. Conf. on Autonomous Agents*, Barcelona.
- Rizzo, P., Veloso, M. V., Miceli, M., and Cesta, A. (1997). Personality-driven social behaviors in believable agents. In *AAAI Symposium on Socially Intelligent Agents*, pages 109–114.
- Rousseau, D. (1996). Personality in computer characters. In *AAAI Technical Report WS-96-03*, pages 38–43.
- Rousseau, D. and Hayes-Roth, B. (1996). Personality in synthetic characters. In *Technical report, KSL 96-21*. Knowledge Systems Laboratory, Stanford University.
- Sansonnet, J. P. and Bouchet, F. (2010). Extraction of agent psychological behaviors from glosses of wordnet personality adjectives. In *Proc. of the 8th European Workshop on Multi-Agent Systems (EUMAS'10)*, Paris, France.
- Silverman, B. G., Cornwell, M., and O'Brien, K. (2006). Human behavior models for agents in simulators and games. In *PRESENCE*, volume 15, pages 139–162.