Subjectivity and Cognitive Biases Modeling for a Realistic and Efficient Assisting Conversational Agent

François Bouchet, Jean-Paul Sansonnet

LIMSI-CNRS
Université Paris-Sud XI

September 16, 2009

IAT’09
Outline

1. Introduction

2. A Subjective and Rational Agent Model

3. Addition of cognitive biases

4. Conclusion
Introduction

 Context: ACA with a cognitive model
 Motivation: improving efficiency through realism

A Subjective and Rational Agent Model

Addition of cognitive biases

Conclusion

François Bouchet, Jean-Paul Sansonnet
Assisting Conversational Agents

Assistance general issues:
- “Paradox of motivation” *(Carroll & Rosson, 1987)*
- Users prefer help from “a friend behind their shoulder” *(Capobianco & Carbonell, 2001)*

ACA seem like an answer:
- “Persona Effect” *(Lester, 1997)*
- Natural Language *(Carbonell, 2003)*

But two believability issues towards realism:
- Physical embodiment
  → going through the “Uncanny valley” *(Mori, 1970)*
- Cognitive abilities
  → improving the human-likeness *(Xuetao et al., 2009)*
Related works

- CoJACK: addition of human physiological constraints to JACK (Evertsz et al., 2008)
- Addition of parameters: fundamental desires, capabilities, resources can help to model emotions (Pereira et al., 2008)
- Order of heuristics: perception of different high level personality traits (Dastani, 2002)
Personality: realism in decisions

“How to quit that application?”

- **Neutral**: “Click on the red button with a cross”
- **Surprise**: “The task isn’t over.” (pragmatics + task context)
- **Sadness**: “You want to leave me?” (past interactions + agent’s subjectivity)
- **Pleasure**: “Good riddance, let me be!” (past interactions + agent’s subjectivity).

Pure rational reasoning isn’t enough:

- Lack of task context = lack of competency
- Lack of subjectivity =
  - lack of realism/human-likeness (user has expectations)
  - lack of coherence (user will interpret it *(Reeves & Nass, 1996)*)
Cognitive constraints: realism in decision-making

**Issues**

- decisions always intentional: the agent can explain them
- emotions don’t have the priority: the agent can inhibit them
  - accidentally: many rules, several designers
  - willingly: if self-monitoring

**Solution**

Special rules $\rightarrow$ *biases*

- hidden: applied outside the agent’s main processing engine
- destructive: the original request can’t be retrieved
Outline

1. Introduction

2. A Subjective and Rational Agent Model
   - Model elements
   - Detailed agent representation
   - Dynamic functioning

3. Addition of cognitive biases

4. Conclusion
Actors

Agent $\mathcal{A}$

$\mathcal{A} = \langle \mathcal{E}, \mathcal{M}, \Psi \rangle$:

- $\mathcal{E}$: set of *agent’s engines*, actively processing requests.
- $\mathcal{M}$: set of *agent’s memories*, storing knowledge of the agent (learnt or original).
- $\Psi$: set of *agent’s mental states*, psychological parameters.

Interacts with the external *world* $\mathcal{W} = \text{users} + \text{application}$. 
**Model elements**

**Information**

\( \mathcal{W}, \mathcal{M} \) and \( \Psi \) store information as *entities*.

**Entity**

Triple associated to an identifier:

\[
#id = H \left[ \bigcup_i a_i \rightarrow v_i \right]
\]

- \( #id \): identifier
- \( H \): head
- \( a_i \): attribute restricted by \( H \)
- \( v_i \): value restricted by \( a_i \): terminal value, other entity, existing identity (identifier)
Communication

- external: $A \leftrightarrow W$
- internal: $E \leftrightarrow M$ and $E \leftrightarrow \Psi$

Handled through messages.

Message

Requests sent between or within actors:

- **INFORM[recipient, request]**: transmits request, expects nothing in return
- **GET[recipient, value]**: asks value, expects an **INFORM[sender,X]** in return
- **CHECK[recipient, attribute, value]**: asks if the value sent is the one of the attribute, expects **INFORM[sender,T|F|?]**
World $\mathcal{W}$

**Definition**

Set of entities providing an “objective” description.

**Information about a user**

```
#user7 = PERSON[
    name   -> "Smith",
    role   -> user,
    age    -> 20,
    gender -> male
]
```
**Definition**

Psychology of the agent, modeled according to four types taking value in \([-1, 1]\) (0 = neutral).

<table>
<thead>
<tr>
<th></th>
<th>Unary</th>
<th>Binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>Trait $\Psi_T$</td>
<td>Role $\Psi_R$</td>
</tr>
<tr>
<td>Dynamic</td>
<td>Mood $\Psi_t$</td>
<td>Relationship $\Psi_r$</td>
</tr>
</tbody>
</table>
Agent’s mental states – Traits $\Psi_T$

**Definition**

Classical “Big Five” (Goldberg, 1981) defining the personality

- **Openness**: appreciation for adventure, curiosity
- **Conscientiousness**: self-discipline and achieves goals
- **Extraversion**: strong positive emotions and sociability
- **Agreeableness**: compassion and cooperativeness
- **Neuroticism**: experience negative emotions easily

**Unary mental state encoding**

```
traits[
    openness -> -0.2,
    conscientiousness -> 0.7,
    ...]
```
Agent's mental states – Moods $\Psi_t$

**Definition**

Personality factors changed in time by heuristics and biases

- **Energy**: physical strength
- **Happiness**: physical contentment regarding the situation
- **Confidence**: cognitive strength
- **Satisfaction**: cognitive contentment regarding the situation
Agent's mental states – Roles $\Psi_R$

**Definition**

Static relationship between the agent and another entity of the world (e.g. users)

- **Authority**: right to be directive to X and reciprocally to not accept directive behaviors from X.
  
  Antisymmetric: $\text{Authority}(X,Y) = -\text{Authority}(Y,X)$

- **Familiarity**: right to use informal behaviors towards X.
  
  Symmetric: $\text{Familiarity}(X,Y) = \text{Familiarity}(Y,X)$

**Binary mental state encoding**

```
roles[
  towards -> #iduser,
  authority -> val1,
  familiarity -> val2]
```
Introduction

A Subjective and Rational Agent Model

Addition of cognitive biases

Conclusion

Detailed agent representation

Agent’s mental states – Relationships $\Psi_r$

Definition

Dynamic relationships between the agent and another entity (e.g. users)

- **Dominance**: power felt towards X.
  Antisymmetric: $\text{Dominance}(X,Y) = -\text{Dominance}(Y,X)$

- **Affection**: attraction and tendency to be nice to X.
  Not necessarily symmetric.

- **Trust**: feeling one can rely on X.
  Not necessarily symmetric.

François Bouchet, Jean-Paul Sansonnet
LIMSI-CNRS Université Paris-Sud XI
Agent's memory $\mathcal{M}$

**Definition**
Stores knowledge learnt through interaction or that the agent originally had.

**Content**
1. Semantic memory $\mathcal{M}_s$: agent’s vision of the world, observed (direct) or created through introspection (indirect).
2. Episodic memory $\mathcal{M}_e$: focused on the agent *i.e.* autobiographical memory (*Tulving, 1983*).
3. Procedural memory $\mathcal{M}_p$: set of heuristics, *i.e.* rules to apply in some given situations, defining the reactions.
Semantic memory $M_s$

**Definition**

Extended subset of the world:

- subset: whole world not available to $A$ and pieces of information possibly out-dated.
- extended: new facts available through reasoning over the memory content.

**World $\mathcal{W}$**

```
#object9 = OBJECT[
    name -> "btnValid2",
    type -> button,
    label -> "OK",
    color -> green
]
```

**Semantic memory $M_s$**

```
#object3 = OBJECT[
    type -> button,
    label -> "OK",
    color -> green,
    trigger-> accept();
]
```
Episodic memory $M_e$

### Definition
Set of previous interactions of the agent with the user and the application, distinguishing incoming (INBOX) from outgoing messages (OUTBOX).

#### INBOX/OUTBOX

**INBOX**

- from -> [sender],
- time -> [timestamp],
- message -> [message]

**OUTBOX**

- to -> [recipient],
- time -> [timestamp],
- message -> [message]
Procedural memory $\mathcal{M}_p$

**Definition**
Set of heuristics defining the reaction to an incoming request.

**Heuristic**
Associates a set of actions to a situation:
- **head**: regular expression defining classes of requests.
- **body**: decision tree, where nodes send messages to $\mathcal{M}$ and $\mathcal{W}$ (rationality) or to $\mathcal{M}_s$ (subjectivity). At the end, an answer request is sent to $\mathcal{W}$. 
Heuristic example

Forbidden action

```plaintext
if conscientiousness > 0 then
    allow ← CHECK[rep, DOABLE[A], true]
end if

if allow = false then
    if agreeableness > 0 then
        if affection(user) ≥ 0 & familiarity(user) ≥ 0 then
            ans ← POS[NOTPOSSIBLE[A]];
        else if affection(user) < −0.5 then
            ans ← NEG[NOTPOSSIBLE[A]];
        else
            ans ← NOTPOSSIBLE[A];
        end if
    end if
else
    ans ← NOTPOSSIBLE[A];
end if

if authority(user) > 0 then
    req ← INFORM[memory, forbidden(A)]
else
    done ← true
end if
```

- sequel -

```plaintext
if neuroticism > 0 then
    req ← INFORM[memory, decrease(satisfaction)]
end if

if dominance(user) > 0 then
    ans ← UNHAPPY
end if

if satisfaction < −0.3 & familiarity(user) > 0 then
    ans ← NEG[(done?ACK:NACK)]
else if done & satisfaction < −0.8 then
    ans ← NEG[(done?ACK:NACK)]
else
    ans ← (done?ACK:NACK)
end if

req ← INFORM[user, answer]
return req
```

Output: [not possible][unhappy][ack/nack]
Engines $\mathcal{E}$

**Natural Language Processing Engine $\mathcal{E}_L$**
- *Grammatical* analysis: lemmatization, POS tagging, WSD...
- *Semantic* analysis: production of a formal request (*Bouchet & Sansonnet, 2007*).

**Behavioral Engine $\mathcal{E}_B$**
- Centralizes the reception and sending of messages
- Chooses heuristics (from $\mathcal{M}_p$) to be applied
- Computes the reactions from heuristics according to current values of $\mathcal{M}_s$ and $\Psi$
Dynamic functioning

Mental States ($\Psi$)

- Traits $\Psi_T$
  - OCEAN
- Roles $\Psi_R$
  - Authority, Familiarity
- Moods $\Psi_t$
  - Confidence, Satisfaction
- Relationships $\Psi_r$
  - Dominance, Affection, Trust

Engines ($\mathcal{E}$)

1. Behavioral Engine ($\mathcal{E}_B$)
2. Natural Language Processing Engine ($\mathcal{E}_L$)
3. Procedural $\mathcal{M}_p$
   - Req-class1 $\rightarrow$
   - if authority(user) > 0 then
   - end if
   - return req
4. Semantic $\mathcal{M}_s$
   - #user7 = PERSON{
     - name -> "Smith",
     - role -> user,
     - age -> 20,
     - gender -> male
   }
5. Episodic $\mathcal{M}_e$
   - Inbox
   - Outbox

Memory ($\mathcal{M}$)

Agent

François Bouchet, Jean-Paul Sansonnet

LIMSI-CNRS Université Paris-Sud XI
Outline

1. Introduction
2. A Subjective and Rational Agent Model
3. Addition of cognitive biases
   - Definition
   - Biases categories and examples
4. Conclusion
Cognitive bias concept

**Definition**

A bias is a transformation rule over messages sent by the agent (within itself or to the world), without the agent’s knowledge. A bias $b$ on a message between $X$ and $Y$: $X \xrightarrow{b} Y$.

**Comparing with heuristics**

- **Objective**: modifying a message
- **Impact**: any request between $X$ and $Y$
- **Factors used**: $\Psi$ only
- **Introspection**: impossible
Representing biases

**Formal definition**

Same structure as heuristics:
- head: category of the bias.
- body: decisions tree to modify the request, according to $\Psi$.

**Biased perception of a nervous and unhappy agent**

```plaintext
BIAS[
  description -> "victimization",
  category -> "perceptive"
  body -> {
    if (neuroticism < -0.5 && satisfaction < -0.9):
      output = NEGATIVE[input]
  }
]
```
### Biases categories

#### Possible channels
- 4 elements ($M, \Psi, \mathcal{E}, \mathcal{W}$) $\Rightarrow$ 6 channels
- Bidirectional channels
- 3 types of messages (INFORM, GET, CHECK)

$\Rightarrow 6 \times 2 \times 3 = 36$ biases possible in theory

#### Restrictions on channels
- $\mathcal{E}_B$ is the core of communication of $A$: it’s the only one able to send messages
- Every message isn’t relevant for each channel
- The agent can always know its mental states $\Psi$

$\Rightarrow 5$ types of biases left on 7 channels
Perceptive bias $\mathcal{W} \xrightarrow{B_p} \mathcal{E}_B$

**Victimization:** cf. previous example

**Minimization:**

*Condition:* $(\text{satisfaction} > 0.5 \land \text{neuroticism} < 0)$

*Consequence:* tend to ignore negativity in user’s NL request.
Expressive bias $\mathcal{E}_B \xrightarrow{B_e} \mathcal{W}$

**Stress:**
*Condition:* authority($A,U$)$< - 0.5$
*Consequence:* extra uncontrollable nervousness in the answer (indepandently from the content of the request).

**Cheeriness/gloominess:**
*Condition:* extraversion$> 0.5$  
(resp. $< - 0.5$)
*Consequence:* adds positive (resp. negatives) connotations to the answer.
Memory retrieval bias $\mathcal{M}_{B_{mr}} \rightarrow \mathcal{E}_B$

(while answering to a GET or CHECK)

**Doubts:**

**Condition:**

\[ \text{trust(agent,agent)} < 0 \land \text{satisfaction} < -0.3 \]

**Consequence:** Discards or lowers the confidence of the facts retrieved from its memory.
Memory access $\mathcal{E}_B \xrightarrow{B_{ma}} \mathcal{M}$

**Bad faith:**

*Condition:* $\text{satisfaction} < -0.8$ && $\text{authority}(U, A) > 0.3$

*Consequence:* Introduce mistakes (e.g. forgetting a parameter) in messages to $\mathcal{M}_S$. Agent is unaware to have done something else than what it was asked for.
Memory storage bias $\mathcal{E}_B \xrightarrow{B_{ms}} \mathcal{M}$

**Tolerance:**
*Condition:* satisfaction $> 0.5$ 
& neuroticism $< 0$

*Consequence:* Do not remember negative comments from the user on the long term (e.g. criticisms towards the agent)

**Scatterbrain:**
*Condition:* conscientious $< -0.3$

*Consequence:* Randomly forget to store some messages said or received into $\mathcal{M}_e$: they are lost forever.
Outline

1. Introduction
2. A Subjective and Rational Agent Model
3. Addition of cognitive biases
4. Conclusion
Summary

- Subjectivity allows to design agents adapting their assistance:
  - a priori: to have a personality $\Psi_T$ matching the user’s one.
  - dynamically: according to the user’s behavior which has modified its mental state ($\Psi_t$ and $\Psi_r$).

- Cognitive biases allow to mimic human cognitive constraints and give primacy to emotions over rationality.
Perspectives

Evaluating novice users interacting with:

1. a purely rational agent
2. a rational and subjective agent
3. a rational, subjective and biased agent

Expected results:

- realism: \(1 < 2 < 3\)
- efficiency: \(1 \leq 2\) and probably \(3 \leq 2\)

But which assisting agent would be the most used? The best rated overall?