Apports Complémentaires de la Subjectivité et des Biais Cognitifs à la Rationalité dans le Contexte de la Fonction d'Assistance

François Bouchet, Jean-Paul Sansonnet

LIMSI-CNRS
Université Paris-Sud XI

June 5, 2009

MFI'09

Plan

Plan

Agents Conversationnels Assistants

Problèmes de l'Assistance :

- "paradoxe de la motivation" *(Carroll & Rosson, 1987)*
- préférence pour "l'ami derrière l'épaule" *(Capobianco & Carbonell, 2001)*

Atouts des ACA :

- "Persona Effect" *(Lester, 1997)*
- Langue Naturelle

. . . mais "Uncanny valley" *(Mori, 1970)* :

- apparence physique
- **"mentale"**

Besoin de personnalité

Raisonnement rationnel parfois insuffisant en termes de :

- *Compétence*: réponse satisfaisant les intentions de l'utilisateur pouvant être liées à la pragmatique linguistique
- *Réalisme*: réaction similaire à un assistant humain et comportement consistant. Crucial pour les émotions *(Ortony, 2003)*.

"Comment je fais pour quitter ?"

- *Surprise* : si la tâche en cours n'est pas achevée (contexte) ;
- *Déception* : s'il apprécie l'utilisateur (subjectivité) ;
- *Satisfaction* : si au contraire l'utilisateur a été impoli (subjectivité).

Besoin de contraintes cognitives

Problème

- si les règles de personnalité sont identiques, il est possible de les supprimer ;
- si plusieurs concepteurs, problème de priorité ;
- si l'agent apprend son comportement optimal ("self-monitoring"), il risque d'éliminer des heuristiques augmentant son réalisme mais diminuant sa compétence.

Conséquences

Nécessité d'avoir des règles cachées et non modifiables.

- appliquées hors du moteur de règles principal ;
- agissant sur les communications internes et externes de l'agent.

Plan

Éléments de base

Entité

Notion à réifier, représentée sous forme de triplet associé à un identifiant :

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

Moteur

Module de génération d'entité E ou de transformation d'une entité E en une entité E'.

Éléments de base

Domaine

Ensemble d'éléments (entités ou moteurs) ayant un accès direct les uns aux autres.

On en distingue quatre :

1. Mémoire de l'agent $\mathcal{M}$
2. États mentaux de l'agent Ψ
3. Moteur de l'agent $\mathcal{E}$
4. Monde $\mathcal{W}$

Représentation du monde $\mathcal{W}$

Définition

Du point de vue de l'agent, tout ce qui n'est pas interne. Il est constitué d'entités évoluant de manière indépendante de l'agent.

Information au sujet d'un utilisateur

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

Mémoire de l'agent $\mathcal{M}$

Définition

Stocke toutes les connaissances que l'agent possède à l'origine ou acquises lors de ses interactions.

Composition

1. Mémoire sémantique : modèle partiel du Monde acquis par consultation (directe) ou inférence (indirecte).
2. Mémoire épisodique : focalisée sur l'agent – mémoire autobiographique *(Tulving, 1983)*.
3. Mémoire procédurale : ensemble d'heuristiques (*i.e.* règles à appliquer selon certaines conditions).

Mémoire sémantique $\mathcal{M}_s$

Définition

Il s'agit d'un sous-ensemble étendu du Monde :

- sous-ensemble : problèmes d'accès et de mise à jour.
- étendu : informations supplémentaires ajoutées par l'application des heuristiques.

Monde

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

Mémoire sémantique

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

Mémoire épisodique $\mathcal{M}_e$

Définition

Ensemble des interactions précédentes de l'agent avec l'utilisateur et l'application. Nous distinguons messages entrants (`INBOX`) et sortants (`OUTBOX`).

INBOX/OUTBOX

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

```
OUTBOX[
   to -> [recipient],
   time -> [timestamp],
   message -> [message]
]
```

Mémoire procédurale $\mathcal{M}_p$

Définition

Ensemble d'heuristiques définissant comment l'agent doit réagir à une requête entrante.

Définition d'une heuristique

Association de réactions à une situation. Une heuristique contient :

- une tête : expression régulière définissant des classes de requêtes.
- un corps : arbre de décision pour construire les réactions de l'agent. Chaque nœud est basé les valeurs retournées par les requêtes envoyées vers $\mathcal{M}$ et $\mathcal{W}$ (rationalité) ou bien vers $\mathcal{M}_s$ (subjectivité). Se termine par un envoi de requête vers le Monde.

Mémoire procédurale $\mathcal{M}_p$

Réaction subjective à une interdiction

"Je te défends d'ouvrir le fichier de configuration !"

```
HEURISTIC[
   description -> "reaction to an interdiction",
   head -> NEG[AUTHORIZATION[
                  granter -> person[id="user"],
                  granted -> person[id="system"],
                  todo -> A___]],
   body -> {
# The action done is acknowledged possibly
# with a negative modalization
# depending on the familiarity,
if (satisfaction < -0.3 && familiarity(user) > 0):
 answer += NEGATIVE[(done?ACK:NACK)]
elif (done && satisfaction < -0.8):
 answer += NEGATIVE[(done?ACK:NACK)]
else:
 answer += (done?ACK:NACK)

# it finally transmits the built answer to the user
INFORM[user, answer]
 }]
```

États mentaux de l'agent Ψ

Définition

Informations concernant la psychologie de l'agent, modélisée selon quatre types de notions (traits, humeurs, rôles et affects) prenant leur valeur dans $[-1, 1]$.

	Unaire	**Binaire**
Statique	Trait Ψ_T	Rôle Ψ_R
Dynamique	Humeur Ψ_t	Affect Ψ_r

États mentaux de l'agent – Traits Ψ_T

Définition

Attributs classiques de la personnalité qui peuvent être considérés comme **stables** au cours de la 'vie' d'un agent – ils correspondent aux "Big Five", couramment utilisés en psychologie *(Goldberg, 1981)* :

- Ouverture,
- caractère Consciencieux,
- Extraversion,
- Agréabilité,
- Neuroticisme.

États mentaux de l'agent – Humeurs Ψ_t

Définition

Facteurs de personnalité qui **varient** avec le temps, en fonction des heuristiques et des biais cognitifs, elles peuvent être physiques ou épistémiques. On distingue :

- *Énergie physique* : force physique de l'agent, au sens large ;
- *Bonheur physiologique* : bien-être physique de l'agent, selon sa situation physique ;
- *Confiance en soi* : force cognitive de l'agent ;
- *Satisfaction (intellectuelle)* : bien-être mental de l'agent, selon l'analyse qu'il fait de sa situation intentionnelle (*i.e.* vis-à-vis de ses B-D-I).

États mentaux de l'agent – Rôles Ψ_R

Définition

Relations **statiques** à caractère institutionnel, entre l'agent et d'autres entités du monde (*e.g.* la relation utilisateur/assistant). On peut définir deux grandes catégories :

- *Autorité* : droit de l'agent à être directif. Cette relation est souvent antisymétrique : `Authority(X,Y) = -Authority(Y,X)`
- *Familiarité* : le droit de l'agent à se comporter de manière informelle. Cette relation est souvent symétrique : `Familiarity(X,Y) = Familiarity(Y,X)`

États mentaux de l'agent – Affects Ψ_r

Définition

Relations **dynamiques** entre l'agent et les autres entités (typiquement les utilisateurs)

- *Dominance* : puissance perçue par rapport à l'interlocuteur. Cette relation est souvent antisymétrique, comme par exemple `Dominance(X,Y) = -Dominance(Y,X)` ;
- *Affection* : attirance et une tendance à agir amicalement de l'agent envers l'interlocuteur. Cette relation n'est pas nécessairement symétrique ;
- *Confiance* : exprime que l'agent fait confiance à l'interlocuteur. Cette relation n'est pas nécessairement symétrique.

Moteurs $\mathcal{E}$

Moteur de Traitement de la Langue Naturelle $\mathcal{E}_L$

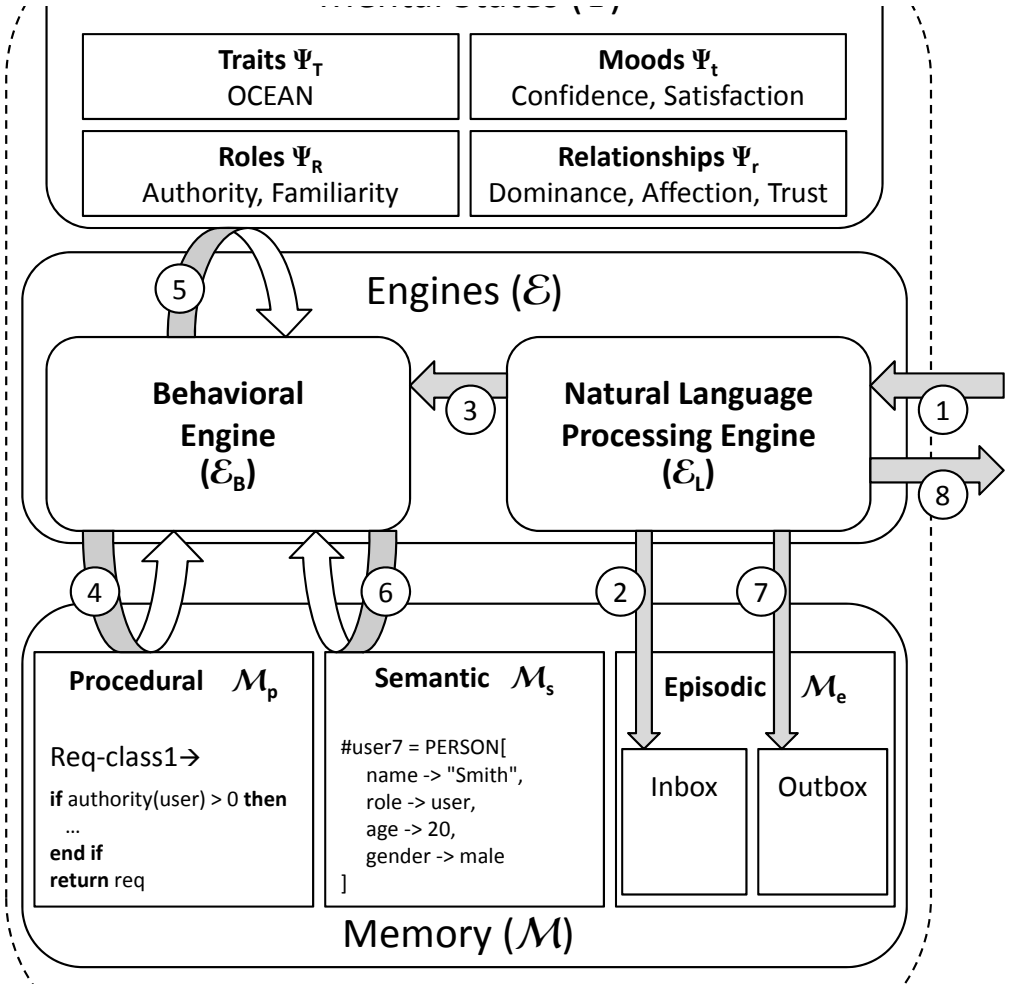
- Analyse **grammaticale** : application d'outils de TALN (lemmatisation, étiquetage, désambigüisation sémantique...) ;
- Analyse **sémantique** : construction d'une requête formelle selon le langage défini dans *(Bouchet & Sansonnet, 2007)*.

Moteur comportemental $\mathcal{E}_B$

Accède par requêtes aux informations des autres domaines :

- `INFORM[domain, request]` ;
- `GET[domain, value]` - attend une requête `INFORM[X,Y]` en réponse ;
- `CHECK[domain, attribute, value]` - attend une requête `INFORM[X,Y]` en réponse (Vrai, Faux, Inconnu).

Fonctionnement dynamique



Plan

Le concept de biais cognitif

Définition

Un biais agit donc comme une transformation sur les requêtes de l'agent sans que celui-ci ne puisse en avoir connaissance.

Biais b sur une requête entre deux domaines X et Y : $X \xrightarrow{b} Y$.

Comparaison avec les heuristiques

- Objectif : **modification** d'une requête
- Portée : **toute** requête entre 2 domaines
- Ψ seulement
- **Non** introspectable

Catégories de biais

Canaux possibles

- 4 domaines $\Rightarrow$ 6 canaux
- Canaux bidirectionnels
- 3 types de requêtes

$\Rightarrow 6 \times 2 \times 3 = 36$ types de biais théoriquement possibles

Hypothèse : restrictions sur les canaux

- $\mathcal{M}$ et Ψ n'ont pas de processus actifs
- Chaque domaine n'envoie pas chaque type de requête
- L'agent connaît toujours ses états mentaux : pas de biais vers Ψ

$\Rightarrow$ **5 types de biais** restant sur 7 canaux

Représentation des biais

Définition formelle

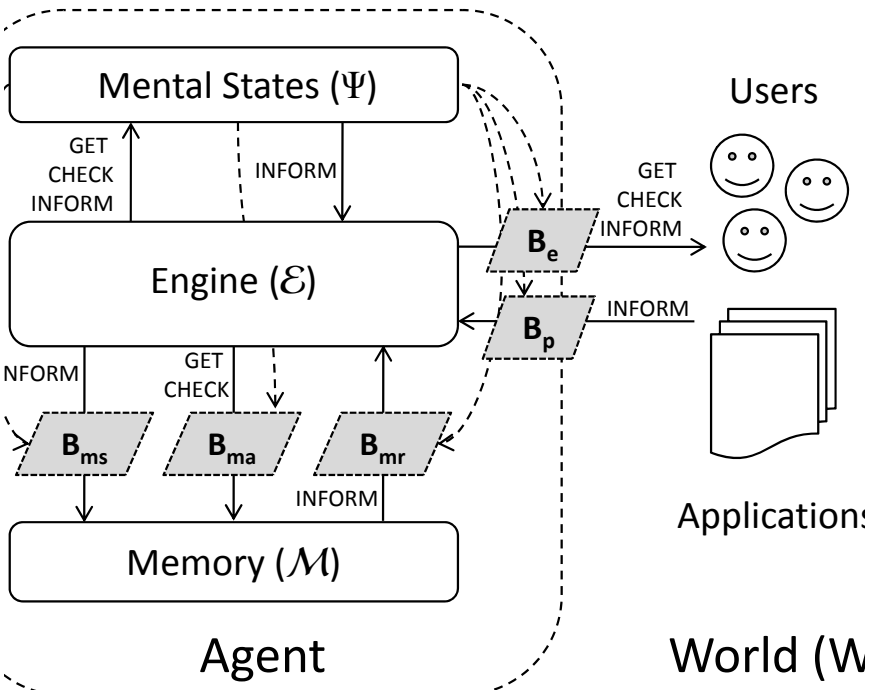
Comme les heuristiques :

- une tête : catégorie du biais parmi les cinq ;
- un corps : arbre de décision pour construire les réactions de l'agent.

Perception par un agent nerveux et malheureux

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

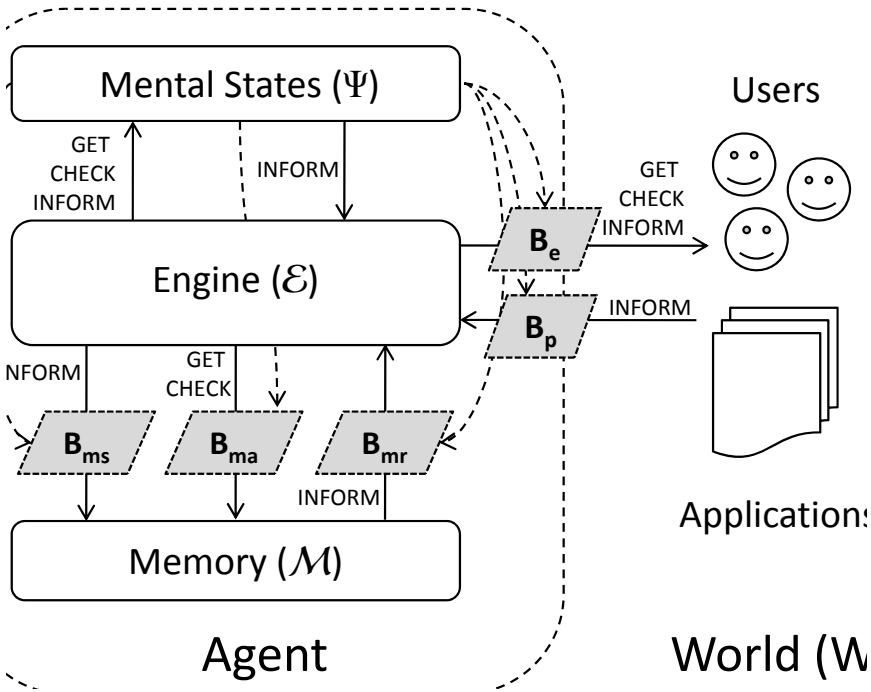
Biais Perceptif $\mathcal{W} \xrightarrow{B_p} \mathcal{E}_B$



Victimisation : *cf.* exemple précédent ;

Minimisation :
Condition : `(satisfaction > 0.5 && neuroticism < 0)`
Conséquence : tendance à sous-évaluer la charge négative contenue dans une phrase de l'interlocuteur.

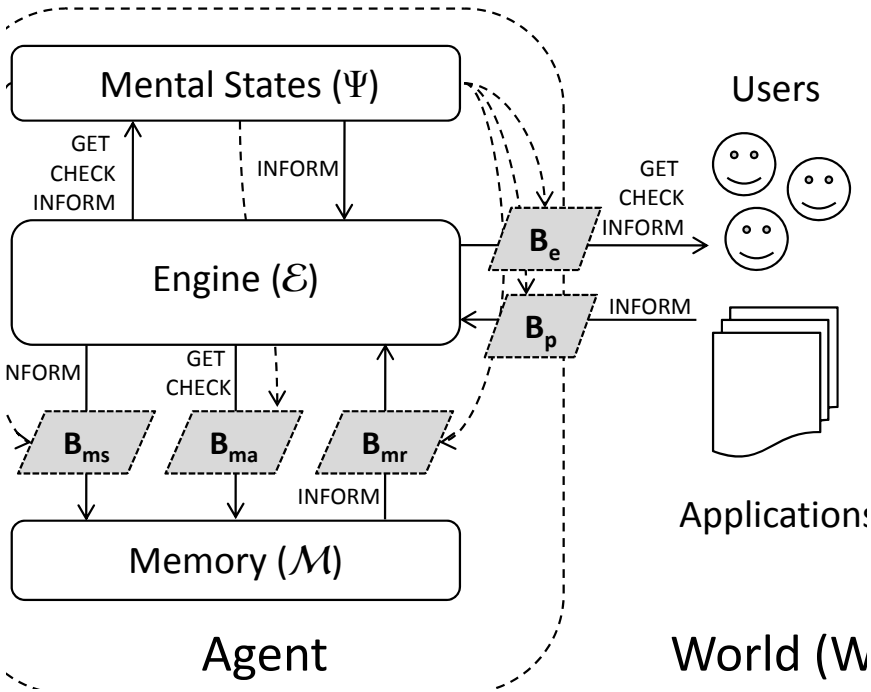
Biais Expressif $\mathcal{E}_B \xrightarrow{B_e} \mathcal{W}$



Stress :
Condition :
`authority(A,U)<-0.5`
Conséquence : nervosité supplémentaire dans ses réactions (indépendamment du contenu propositionnel de la requête).

Enjouement/tristesse :
Condition : `extraversion>0.5` (resp. `<-0.5`)
Conséquence : ajout de connotations positives (resp. négatives) à ses réponses.

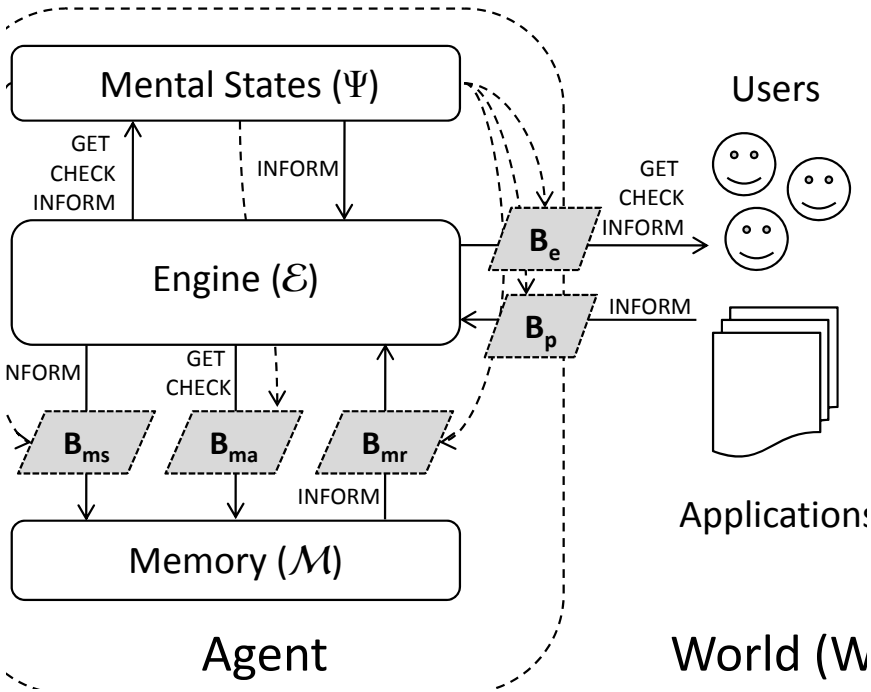
Biais de Recherche $\mathcal{M} \xrightarrow{B_{mr}} \mathcal{E}_B$



(réponse à `GET` ou `CHECK`)

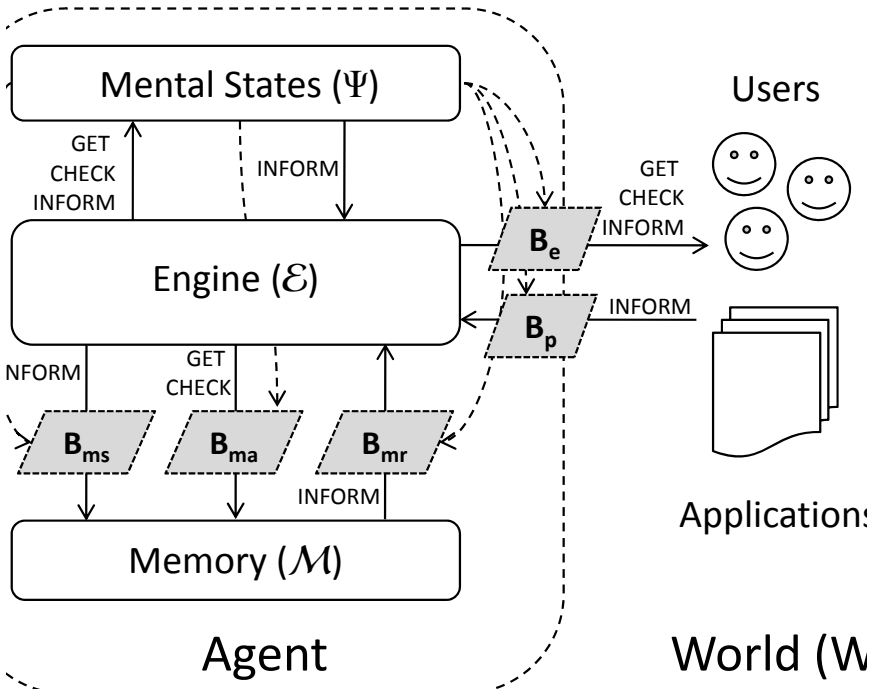
Doute :
Condition :
`trust(agent,agent)<0 && satisfaction<-0.3`
Conséquence : Rejet ou minimisation des informations issues de sa mémoire.

Biais de Lecture $\mathcal{E}_B \xrightarrow{B_{ma}} \mathcal{M}$



Mauvaise foi :
Condition : `satisfaction<-0.8 && authority(U,A)>0.3`
Conséquence : Introduction de fausses informations dans les requêtes vers $\mathcal{M}_s$(ex : oubli d'un paramètre $\Longleftrightarrow$ "acte manqué")

Biais d'Écriture $\mathcal{E}_B \xrightarrow{B_{ms}} \mathcal{M}$



Oubli :
Condition : `satisfaction>0.5 && neuroticism<0`
Conséquence : Ne pas mémoriser certaines informations négatives (ex : critique de l'utilisateur).

Désordonné :
Condition : `conscientious<-0.3`
Conséquence : Perte au hasard d'une part du contenu propositionnel de la requête vers $\mathcal{M}_e$.

Plan

Conclusion

- Prise en compte de la subjectivité permet d'adapter, de manière générique, l'aide à l'utilisateur :
  - statiquement : en adaptant les traits de personnalité de l'agent pour qu'ils soient similaires à ceux de l'utilisateur.
  - dynamiquement : en fonction du feedback de l'utilisateur.
- Les biais cognitifs, ne pouvant être introspectés, donnent une primauté aux états mentaux sur le raisonnement rationnel et permettent modélisation de comportements complexes.
- Architecture implémentée à la chaîne de traitement de la langue naturelle.

Perspectives

Évaluation d'usagers novices face à trois classes d'agents :

- un agent purement rationnel ;
- un agent rationnel et subjectif sans biais ;
- un agent rationnel et subjectif, avec biais.