Intelligence artificielle et conscience

Gérard Sabah

GT « *vers une technologie de la conscience ?* »

La seule façon d'exister pour la ***conscience*** *c'est d'avoir* ***conscience*** *qu'elle existe.*

Jean-Paul Sartre

ACADÉMIE DES TECHNOLOGIES

POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

Menu

- Généralités IA classique, ia « actuelle »
- Conscience, de quoi parle-t-on ?
- Revue rapide de quelques auteurs
- Examens détaillé de quelques théories
  - *Baars*, *Edelman*, *Cardon*, *Pitrat*, *Sabah*
- Conclusion

Ce qu'on appelle « intelligence » n'est pas une faculté unique, mais un ensemble de compétences, innées ou acquises. - Shutterstock

Il n'y aura jamais d'intelligence artificielle

Luc de Brabandere / Boston Consulting Group | Le 23/11 à 07:00, mis à jour à 16:23

L'idée d'intelligence artificielle présuppose qu'il n'existe qu'une seule forme d'intelligence. C'est bien sûr loin d'être le cas...

Caractériser l'intelligence ?

- Capacités de calcul efficaces
- Résolution de problèmes
- Mémorisation et accès aux connaissances
- Catégorisation des observations
- Capacité d'adaptation aux situations imprévues
- Réflexivité et auto-évaluation
- …

Principaux outils de l'IA « classique »

- *Arborescences, min-max, αβ…*
- *Systèmes experts*
- *Logiques non classiques (floues, modales, non monotones)*
- *Programmation par contraintes*
- *Réseaux de neurones*
- *Algorithmes génétiques*
- *Raisonnement par analogie, à partir de cas*
- *Les systèmes multi-agents et l'IA distribuée*

QUATRE NIVEAUX D' « INTELLIGENCE »

- 1) Programme donné à l'avance
  - *Les informations pour l'exécution sont connues a priori*
- 2) Programme adaptable selon l'environnement
  - *Les informations sont cherchées dans l'environnement*
- 3) Programme adaptable par apprentissage
  - *Progression vers l'autonomie*
- 4) Système de systèmes reconfigurable
  - *Systèmes multi-agents massifs non indépendants*
  - *Validation ?*

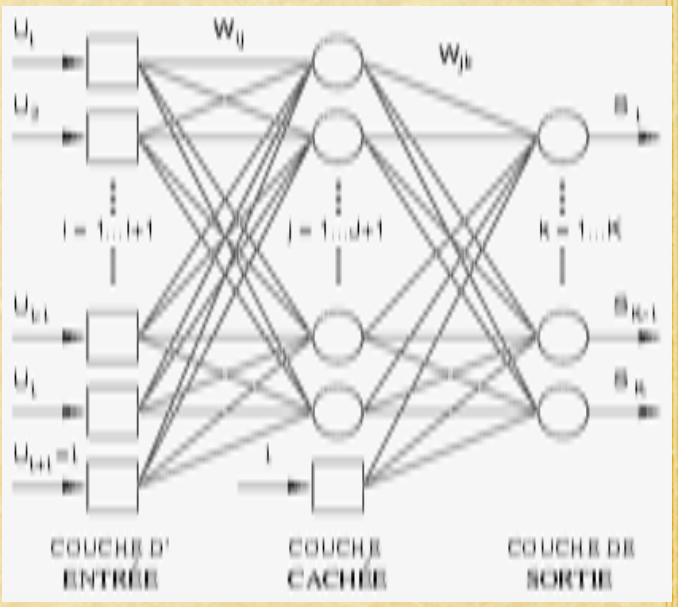
À la mode actuellement :

« Deep learning et big data »

Apprentissage profond
lié aux données massives

Réseaux de neurones artificiels

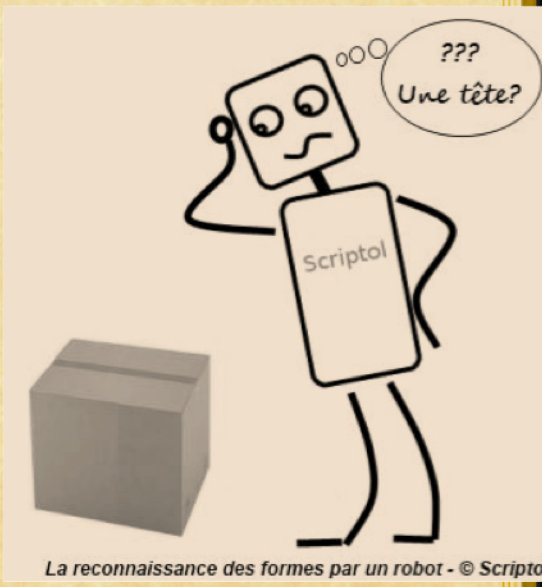
- Perceptron (Rosenblatt 1957)
  - ≠ *Minsky & Papert (1969)*
- Réseaux multicouches
  - *Rumelhart, Yann LeCun*
  - *Convexité nécessaire*
- Apprentissage profond
  - *Apprentissage hiérarchique de concepts intermédiaires*
  - *Utilisation de GPU (Graphic Processing Unit)*



Apprentissage profond et données massives

- Plusieurs dizaines de couches intermédiaires
- → Caractéristiques intermédiaires abstraites. MAIS
- Pas de réelle compréhension ni de possibilité d'explication
- Ni réflexivité ni « conscience »

vs. apprentissage/1-3 exemples



La reconnaissance des formes par un robot - © Scriptol

Complexité biologique

- Cerveau, : ~ 10^{10} neurones dans le cortex, 10^{15} connexions, 10^{9} connexions par mm^3. Combinaisons possibles d'états du cerveau : 10plusieurs millions

- Méthodes scientifiques non adaptées à l'étude du vivant (complexité psychologique des observateurs)

- Caractéristique de l'esprit : capacité à *faire référence*
  - ✚ le cerveau change continûment ses caractéristiques, selon ses relations avec le monde : il est totalement impossible de retrouver à l'identique un état antérieur.
  - ⇨ Impossibilité de représenter l'ensemble des états des neurones (il y faudrait 100 milliards de bits par seconde !)

Pourquoi l'IA s'intéresse-t-elle à la conscience ?

- Edelman : *Les fonctionnalités nécessaires à une véritable intelligence sont celles qui, fondées sur l'inconscient, permettent l'émergence de la conscience chez l'homme*

- *Pourquoi pas chez les machines ?*

Définition(s) [Robert]

- Usage courant
  - *être éveillé (se rendre compte)* — ***awakeness***
  - *connaissance immédiate (spontanée)* — ***awareness***
- Conscience psychologique
  - *ce que ça fait « d'être X »* — ***awareness / consciousness***
  - *connaissance de sa propre activité psychique et de ses opérations mentales (introspection)* — ***consciousness***
- Définition physique (mécanique quantique ?)
- Conscience morale — ***conscience***
  - *faculté de juger ses propres actes (Bien, Mal)*

Quatre caractéristiques

- *Sélectivité*
  - *Tout n'arrive pas à la conscience*
  - *Fonction de sélection de la conscience*
- *Exclusivité*
  - *Effet de séquentialisation*
  - *Tous les niveaux ne sont pas conscients*
- *Enchaînement*
  - *Les événements conscients sont traités en série*
  - *Fonction constructive de la conscience*
- *Unité*
  - *Ce qui fait que l'esprit est un tout*
  - *Perceptions et réalité ?*

Différences entre processus

Conscients	Inconscients
Calculs inefficaces (erreurs, lents, interférences avec autres processus conscients)	Efficaces pour leur tâche spécifique (peu d'erreurs, rapides, pas d'interférences)
Contenus très variés, liens avec processus conscients et contextes inconscients	Domaine limité, strictement défini, autonomes
Volumes limités, sériels, cohérence interne	Traitent de grands volumes, très divers, opèrent en parallèle sans interactions

Où se situe « la » conscience ?

- En fin de traitement (processus d'interprétation)
- Au début des processus de traitement où les résultats des traitements sensoriels forment encore un tout et préservent les relations spatiales de la scène originelle
- Sur l'ensemble du cerveau qui joue le rôle d'un observateur de ces propres résultats et influe sur eux pour maximiser la reconnaissance

Hypothèses

- Hypothèse de l'activation
  - *Ce n'est pas l'activité elle-même.*
  - *Perte de conscience des événements répétés et prédictibles ?*
  - *~ Probabilité d'un événement à devenir conscient*
- Hypothèse de la nouveauté
  - *Lié à la notion d'information*
  - *Deviennent conscients les éléments apportant de l'information.*
- Hypothèse du sommet de l'iceberg
  - *Conscient = émergence d'expériences inconscientes*
  - *Limites de la conscience ?*
- Hypothèse du théâtre
  - *Conscience vue comme un lieu dans le cerveau où les informations produites par les traitements issus de nos sens sont collectées pour être rendues conscientes*

Confusions à éviter

- « *conscience*, *attention*, *expérience*… » utilisés comme si
  - sens clair et univoque
  - division claire entre les choses auxquelles ils s'appliquent ou ne s'appliquent pas
- D'où diverses « pseudo-questions »
  - *quels animaux sont conscients ?*
  - *comment la conscience évolue-t-elle ?*
  - *quelle est sa fonction biologique ?*
  - *est-on conscient d'un bruit si on ne s'en rend compte qu'au moment où il s'arrête ?*
  - *un robot peut-il être conscient ?*
  - *une machine peut-elle avoir les apparences externes de la conscience sans être consciente ?*

Ni dichotomie, ni continuum

- Pour tenter de répondre aux questions précédentes :
  - *conscience = question de degré*
  - *conscience = faisceau de concepts, liés aux capacités :*
    - de perception
    - d'attention (implicite / explicite)
    - de mémoire(s)
    - de langage (*rapportabilité*)
    - d'apprentissage (explicite et implicite)
    - d'évaluation et contrôle de soi
    - des émotions
    - de représentation d'états mentaux d'autrui
    - …

Quelle est la bonne question ?

- Un robot ou un système est-il (ou pourra-t-il être) *conscient* ?
  Peut-on reproduire la conscience humaine dans une machine ?

- Quelles sont les fonctionnalités que l'homme attribue à sa conscience et qui pourront être mises en œuvre dans les futurs robots ?

Fonctionnalités

- Interprétation et unification des données sensorielles
- Représentation et interprétation autocentrées de l'environnement
- Reconstruction de scènes passées
- Constructions imaginaires
- Perception et représentation de soi
- Choix de comportements appropriés
- Gestion des intentions et planification
- Gestion de l'attention
- Réactions face à des événements imprévus
- Résolution de problèmes, planification des buts
- Gestion des hypothèses
- Déclenchement de processus planifiés
- Contrôle de la réalisation des objectifs ; replanification
- Gestion de la mémoire à court terme
- Mémorisation des événements
- Apprentissage
- Rendre l'information pertinente accessible
- Décision d'apprentissage volontaire
- Attribuer les désirs, les intentions, les motivations à d'autres
- Se représenter autrui
- Prédire les comportements (de soi ou des autres)
- Permettre une communication sociale efficace
- Intériorisation des règles sociales
- Constitution d'une personnalité constante

Quelques points de vue synthétiques sur la conscience

Ceux marqués d'une * = présentation
plus détaillée possible

Auteurs pertinents

- Edelman*
- Baars
- Harth
- Cardon*
- Chalmers
- Damasio
- Dennett
- Eccles
- Husserl
- Jeannerod
- Jacquendoff
- Johnson-Laird
- Maturana et Varela
- Minsky
- Ornstein
- Penrose
- Pitrat*
- Rosenfield
- Sabah*

Edelman (*)

- Conscience fondée sur une théorie des fonctions du cerveau, fondée elle-même sur l'évolution et le développement.
- TSGN (Théorie de la Sélection des Groupes de Neurones) ;
  3 principes :
  - *sélection ontogénétique* ;
  - *renforcements ou affaiblissements synaptiques secondaires* ;
  - *interaction de cartes cérébrales avec réentrance*.

Baars (*)

« La conscience est au psychologue ce que la gravité est au physicien : inévitable »

- Caractéristiques d'une expérience consciente :
  - *implique une diffusion globale de l'information*
  - *implique une cohérence interne (≠ du rêve)*
  - *demande à être adaptée au reste du système*
  - *demande accès par le système du soi*
  - *peut nécessiter des perceptions d'une certaine durée*
- Conception « économique » de la conscience (seuls trois concepts essentiels) :
  - *une zone de travail globale,*
  - *des processus inconscients spécialisés*
  - *des contextes (hiérarchies de buts)*

Harth

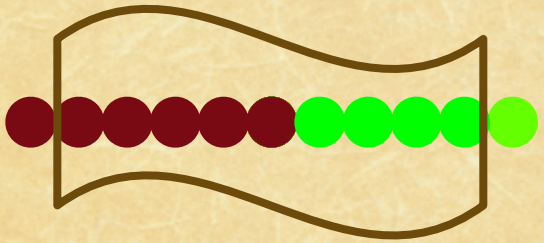
- Rejette aussi bien le dualisme cartésien que le pluralisme radical
- Auto-référence et boucles de rétroaction : les images mentales sont combinées avec les connaissances antérieures
- « Boucles créatrices »
  - *Pas d'homoncule qui examine l'état du cerveau, c'est le cerveau qui analyse et recrée, puis observe à nouveau ses créations, visant à maximiser la reconnaissance*
- Chemins « descendants » fondamentaux
- Cohérent avec les rétroactions d'Edelman
- Ce mécanisme apparaît au début du réseau sensoriel

Harth : boucles créatrices

- Ces boucles de rétroaction sont la règle dans l'ensemble des mécanismes du système nerveux
- Mécanisme instable (amorçage ; permet l'adaptation dans un environnement changeant)
- Bruits aléatoires : permet de sortir de minimums locaux (*une explication possible de la créativité*)
- **Les perceptions modifient les connaissances ET les connaissances modifient les résultats des perceptions** ***(Staline vs Orwell)***

Rétroactions (niveau subliminal)

- Interactions de différents niveaux
  - *Cf. vision, (chemins « descendants »)*
  - *modification active du message purement visuel (injection d''informations supplémentaires)*
- Images mentales ≠ répliques des objets du monde ; systématiquement combinées avec les connaissances antérieures



Cardon (*)

- Une pensée artificielle est calculable ; nécessité d'un corps matériel
- Composants élémentaires : actifs, proactifs, symboliques, évolutifs et communicants ; regroupés en *agents aspectuels* (contexte). Fonctionnement réflexe
- Des *agents morphologiques* représentent l'état et le fonctionnement des agents aspectuels. Co-activité et influence réciproque des deux systèmes (« méta »)
- SMA massivement parallèle
- Stabilisation de l'ensemble = « état de pensée »

Cardon (ste)

- *Émotion*
  - *non consciente, comportementale, automatique*
- *Sentiment*
  - *conscient, déclenché par une émotion*
- connaissance d'un *sentiment*
  - *1re approche de la conscience (réflexive)*
- calcul d'un *plan d'action*
  - *raisonnement*
- *conscience noyau*
  - *perception de soi-même en état d'appréhension du monde)*
- *conscience étendue*
  - *mise en situation de l'individu dans la temporalité*

Chalmers

- Inventeur de l'expression « *hard problem* »
  (pourquoi ∃? expériences phénoménales qualitatives)
  - *pourquoi l'activité de notre cerveau nous fait-il ressentir quelque chose plutôt que rien ?*
- Double analyse : physique et phénoménale
- Concilier divers points de vue
  - *Fonctionnalistes (bio et IA)*
  - *Neurobiologistes (structuralisme)*
  - *Phénoménologistes (qualia)*

Damasio

- Importance des émotions pour le raisonnement et la conscience
- Protosoi
  - *Cartes cérébrales + images mentales du corps (boucle)*
  - *Le corps* ***est le fondement de l'esprit conscient***
- Soi-noyau
  - *Relations avec les objets extérieurs (inclut les sentiments)*
  - *Base de l'esprit conscient*
- Soi autobiographique
  - *Émerge de l'agrégation des multiples images du soi-noyau*

Dennett

- Pandémonium de « versions multiples »
  - *mécanisme permettant le « recrutement » des processeurs d'interprétations et d'élaborations activés en parallèle*
- Seuls les résultats jugés pertinents au contexte sont « rendus conscients »
  - *pertinence = phénomène un peu "magique" (cf. heuristiques de l'intelligence artificielle)*
- Un effet de diffusion générale, (isotropie) permet à ce qui est conscient d'influencer n'importe quoi
  - *Conscience = machine virtuelle sur le matériel parallèle du cerveau.*
- Rôle essentiel du langage
  - *le langage infléchit nos pensées à tous les niveaux. […] Les structures de la grammaire imposent une discipline à nos pensées…"*

Eccles

- Se fonde sur la notion de dualisme
- Distingue trois mondes (relations récursives)
  1. *matériel*
  2. *états de conscience (pensées subjectives)*
  3. *connaissances objectives*
     - processus de rétroaction entre le monde 2 et le monde 3
- Hypothèse des micro-sites (physique quantique)
  - *événements mentaux agissent sur événements neuraux*
  - *esprit est comparé à un champ de probabilités quantique permettant l'activation de vésicules du réseau présynaptique*
  - *L'esprit conscient : lecture passive des opérations du cerveau + activité propre de recherche + un rôle d'unification de l'ensemble*

Husserl

- Fondateur de la phénoménologie
- Étude de phénomènes, fondée sur l'analyse de l'expérience vécue par un sujet
- Toute conscience doit être conçue comme « conscience de quelque chose »
- Modes d'accès de la conscience à la signification
  - *conscience définie par l'**intentionnalité** : le sujet est toujours lié, corrélé à l'objet qu'il perçoit, imagine, etc.*

Jackendoff

- Perception, action, pensée et apprentissage sont tous inconscients
- Contenus conscients = entités intermédiaires
- Conscience censée contenir les distinctions essentielles entre les choses

[*sans qu'il donne la moindre idée de pourquoi ni comment cela peut se passer comme ça…*].

Johnson-Laird

- Conscience analogue à un système d'exploitation informatique.
- Moniteur de haut niveau gère l'ensemble des processus inconscients qui agissent en parallèle
  - *Assigne des priorités relatives aux processus en attente*
  - *Gère des interruptions par l'intermédiaire de systèmes de sémaphores complexes*
- Conscience = mode de fonctionnement particulier de ce système, disposant d'un modèle de lui-même.
  - *[Idée assez extraordinaire …].*

Maturana et Varela

- *énaction* : cognition en relation avec le corps et l'environnement (la cognition incarnée)
- Prise en considération du boudhisme et de la phénoménologie (*expérience personnelle à l'origine de la connaissance*)
- Autopoïése : réseau de processus de production de composants
  - *régénèrent continuellement par leurs transformations et leurs interactions le réseau qui les a produits*
  - *constituent le système en tant qu'unité concrète dans l'espace où il existe, en spécifiant le domaine où il se réalise*

Minsky

- La « société de l'esprit » ≠ modèle de la conscience,
- La conscience ne concerne que le passé (non le présent)
- La connaissance est atomisée en un ensemble (gigantesque) d'agents très simples qui interagissent
- N'explique pas le rôle de la conscience dans la constitution de l'objectivité ni dans les relations au monde extérieur.
- Probablement une origine des idées de Dennett

Ornstein

- Les capacités mentales de haut niveau sont accidentelles (*serendipity*)
- Simpletons :
  - *capacités inconscientes du cerveau*
  - *des « escadrons » passent d'un état contrôlé à un état non contrôlé*
  - *Les modifications et les évolutions conscientes du comportement passent alors par une compréhension profonde et une gestion de ce processus*
- Théorie proche de la « société de l'esprit » de Minsky (qui d'ailleurs n'est pas cité)

Penrose

- Théorème d'incomplétude de Gödel → la conscience ne peut être réduite au fonctionnement d'un système formel
- Mécanique quantique à la base de la conscience
- Il argumente pour que ce type de décision soit du ressort de mécanismes réflexifs et en particulier de la conscience
- Diverses failles dans ce raisonnement

Rosenfield

- Construction perpétuelle du ***sens*** comme relation entre les catégorisations perceptuelles, les émotions et les expériences physiques et sociales
- Construction d'une image cohérente du corps et de ses mouvements qui aboutit à une image de Soi
- Introduction des niveaux perceptuel, conceptuel et linguistique.
- **Conscience vue comme un processus reliant nos souvenirs à la représentation actuelle de nous-mêmes (*importance du langage*)**

Points de vue détaillés

Gerald Edelman

Alain Cardon

Jacques Pitrat

Gérard Sabah

Convergences

- Résolution de problèmes (J.P.)
  - → *conscience réflexive + conscience morale*
- Systèmes multi-agent
  - *Stabilité→ état de pensée (A.C.)*
- Traitement automatique des langues (G.S.)
  - *Inspirations d'Edelman, Baars et Harth*
  - → *inconscient + conscience réflexive*

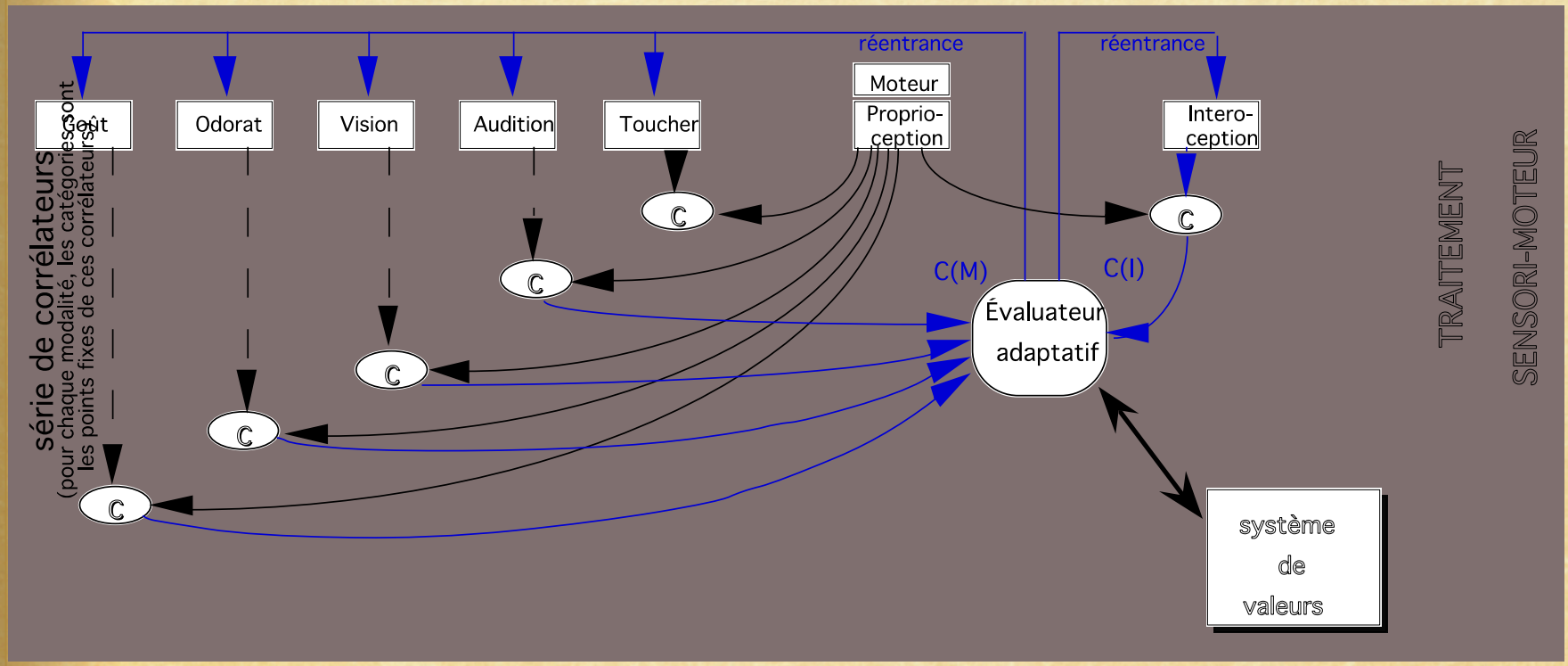
Edelman

- Conscience fondée sur une théorie des fonctions du cerveau, fondée elle-même sur l'évolution et le développement.
- TSGN (Théorie de la Sélection des Groupes de Neurones) ; 3 principes :
  - *sélection ontogénétique* ;
  - *renforcements ou affaiblissements synaptiques secondaires* ;
  - *interaction de cartes cérébrales avec réentrance*
    → *rôle essentiel de la catégorisation*

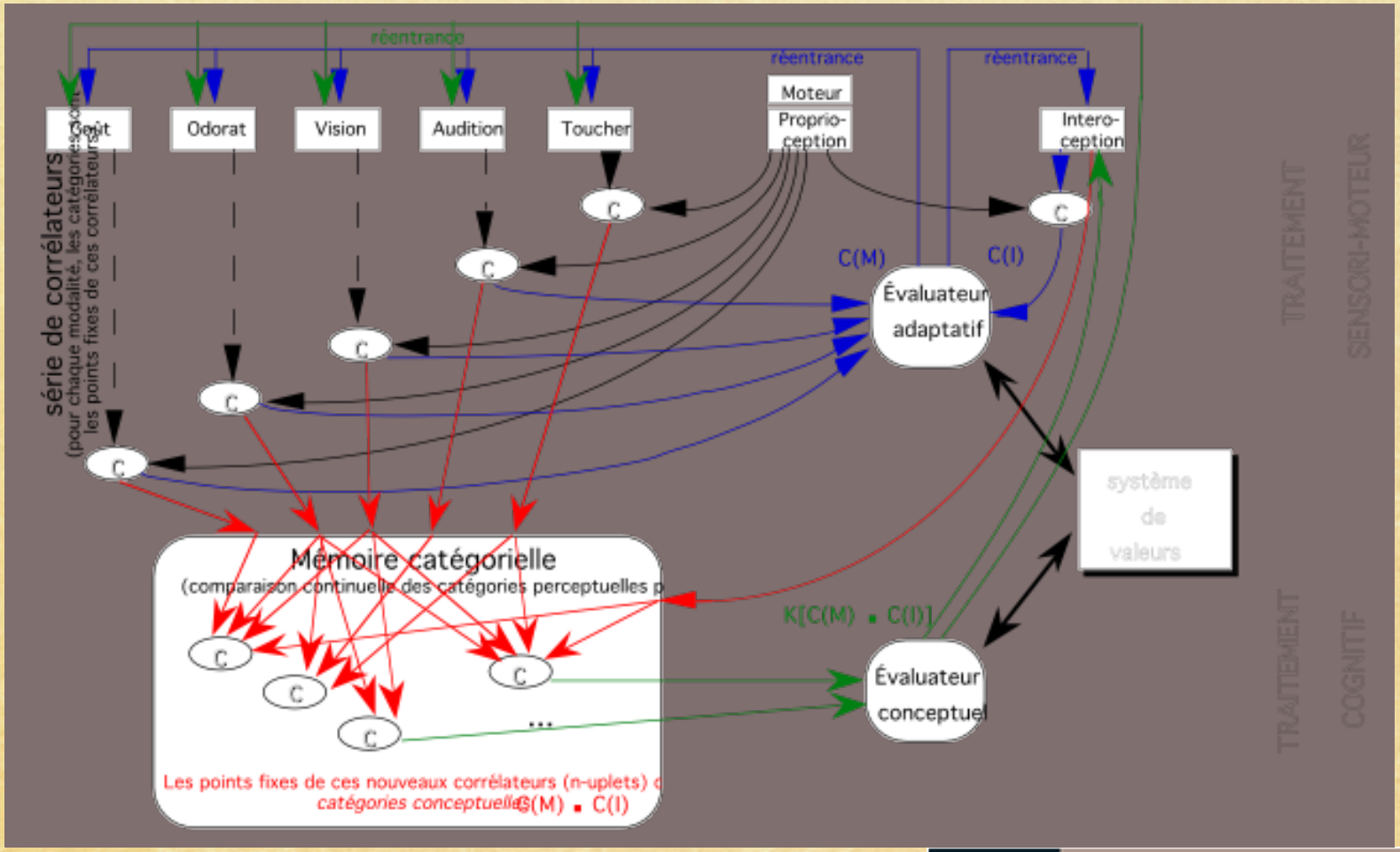
Edelman

- Fonctions du niveau neurobiologique :
  1. *spécialisations neurales permettant de distinguer les signaux internes des signaux du monde*
  2. *catégorisation perceptuelle*
  3. *mémoire (s) = processus de recatégorisation continuelle*
  4. *apprentissage (liens catégories ↔ valeurs)*
  5. *acquisition de concepts*
  6. *conscience primaire*
  7. *capacité d'ordonnancement, (présyntaxe, base des capacités symboliques)*
  8. *langage*
  9. *conscience d'ordre supérieur.*

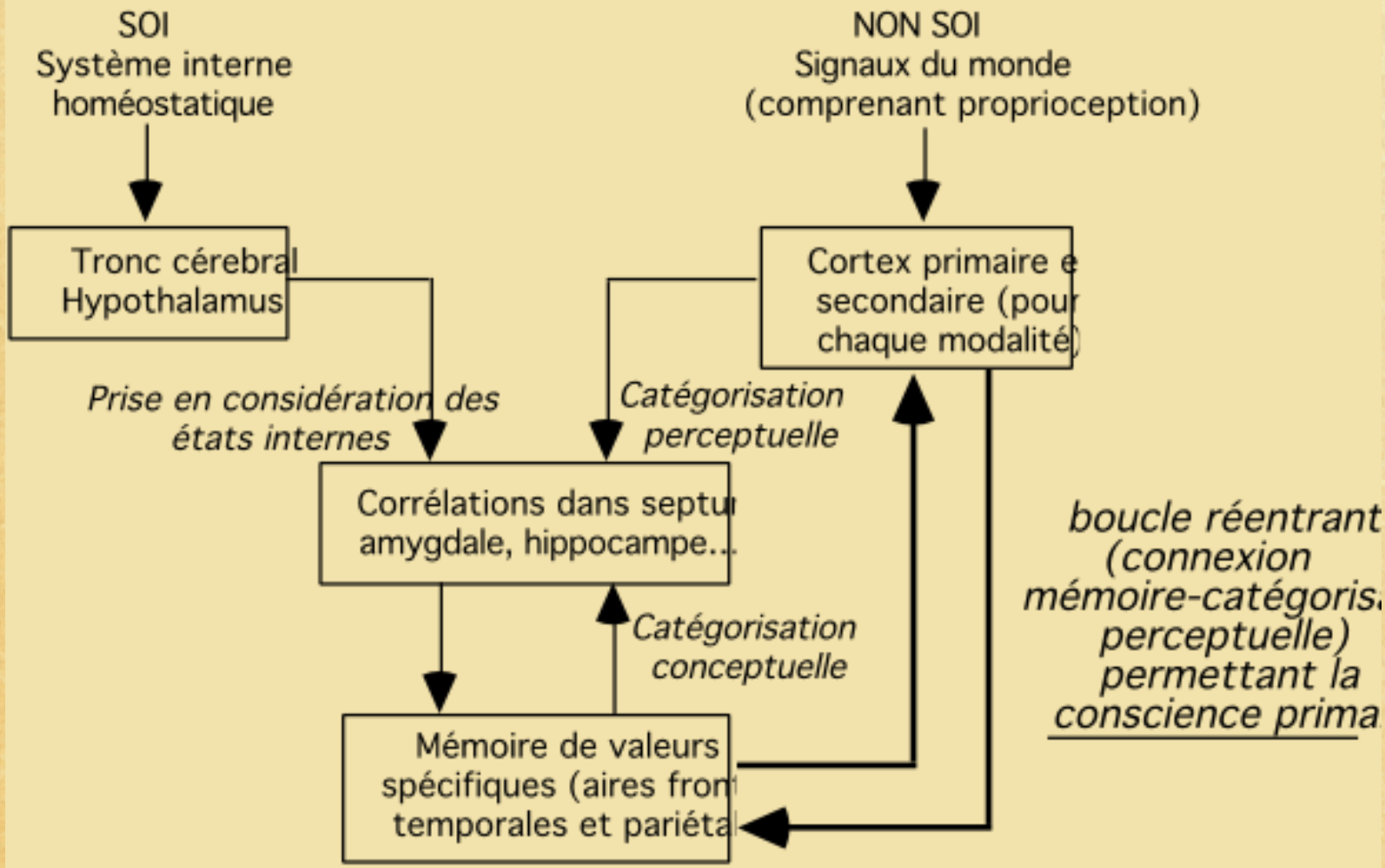
Edelman niveau sensori-moteur



Edelman niveau cognitif



La conscience primaire



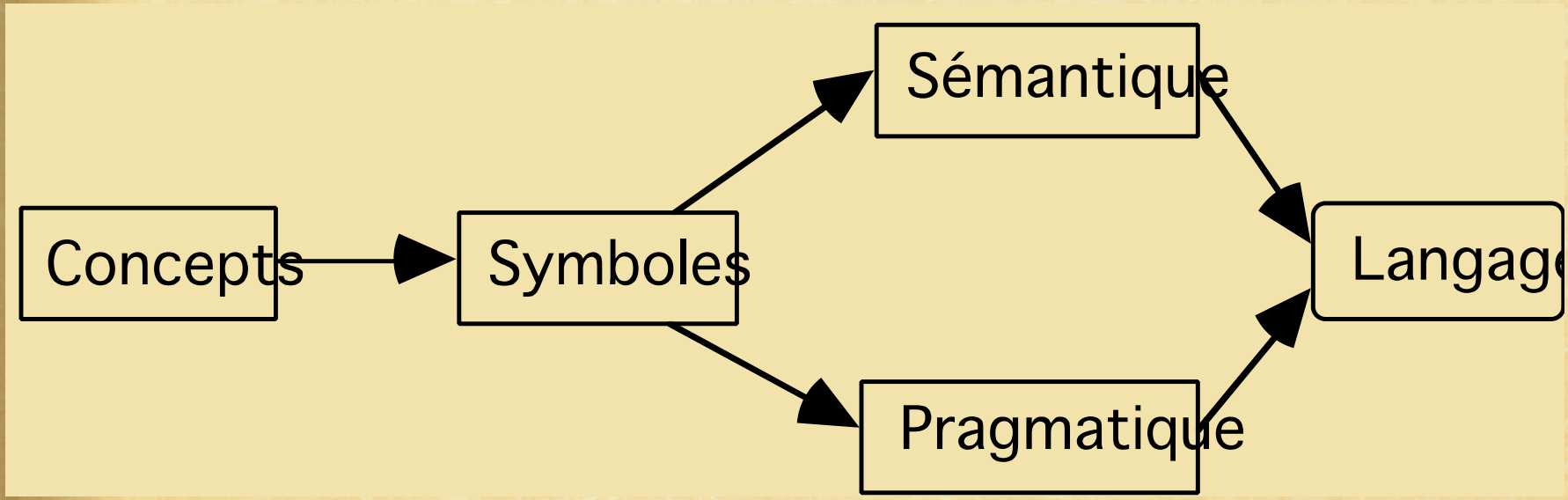
Conscience primaire (ste)

- Capacité à créer des *scènes*
- Comparaison discriminative entre mémoire conceptuelle et catégorisations perceptuelles
  → conscience primaire des objets et des événements
- Altération dynamique de la mémoire
- Pas de régression infinie

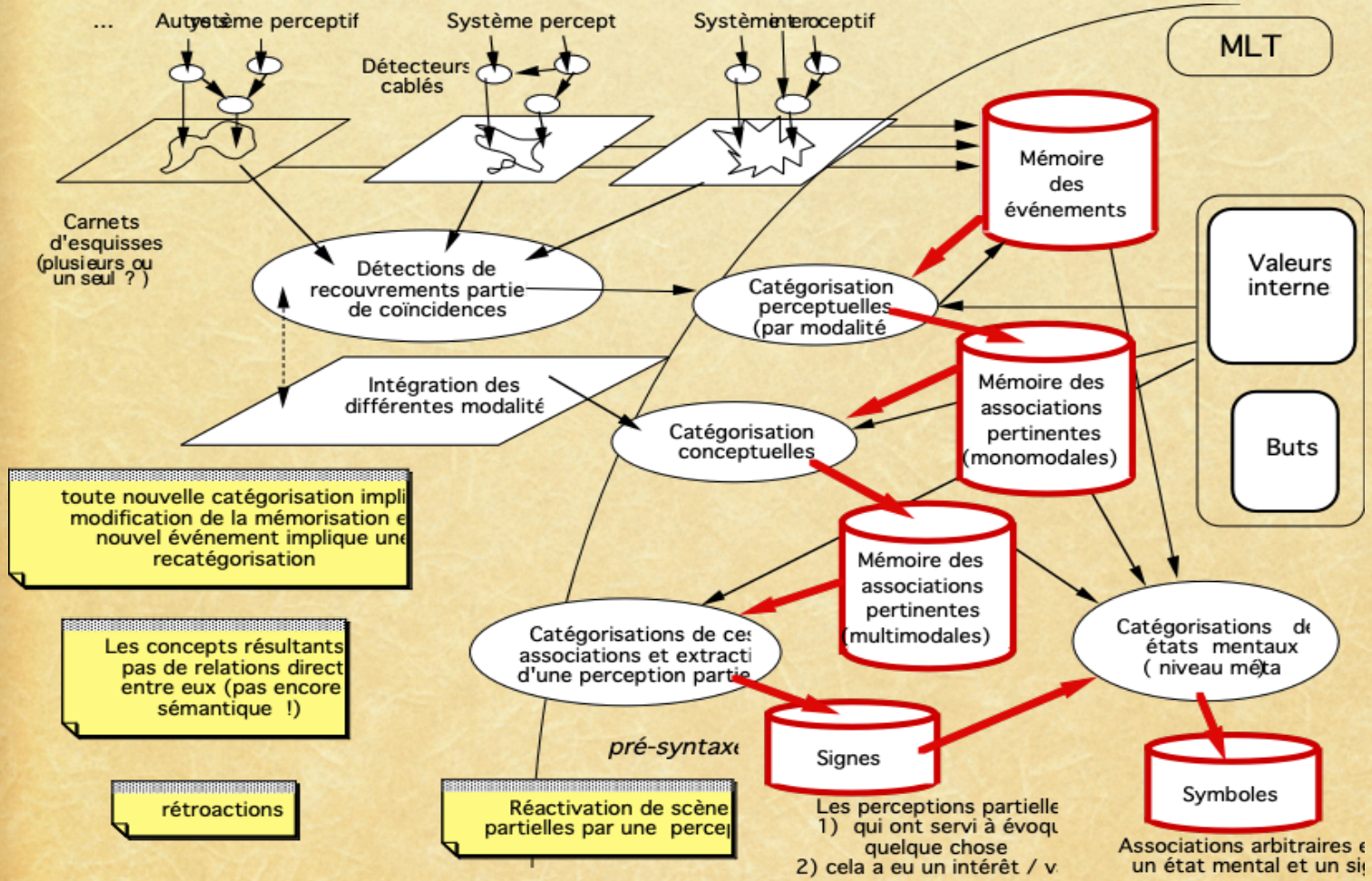
Ordonnancement

- La formation de concepts ne suffit pas à expliquer la pensée
- Nouveau type de mémoire → *concepts dans une relation ordonnée*
- reconstruction partielle de *scènes* → présyntaxe → capacité de s'abstraire des contraintes du présent

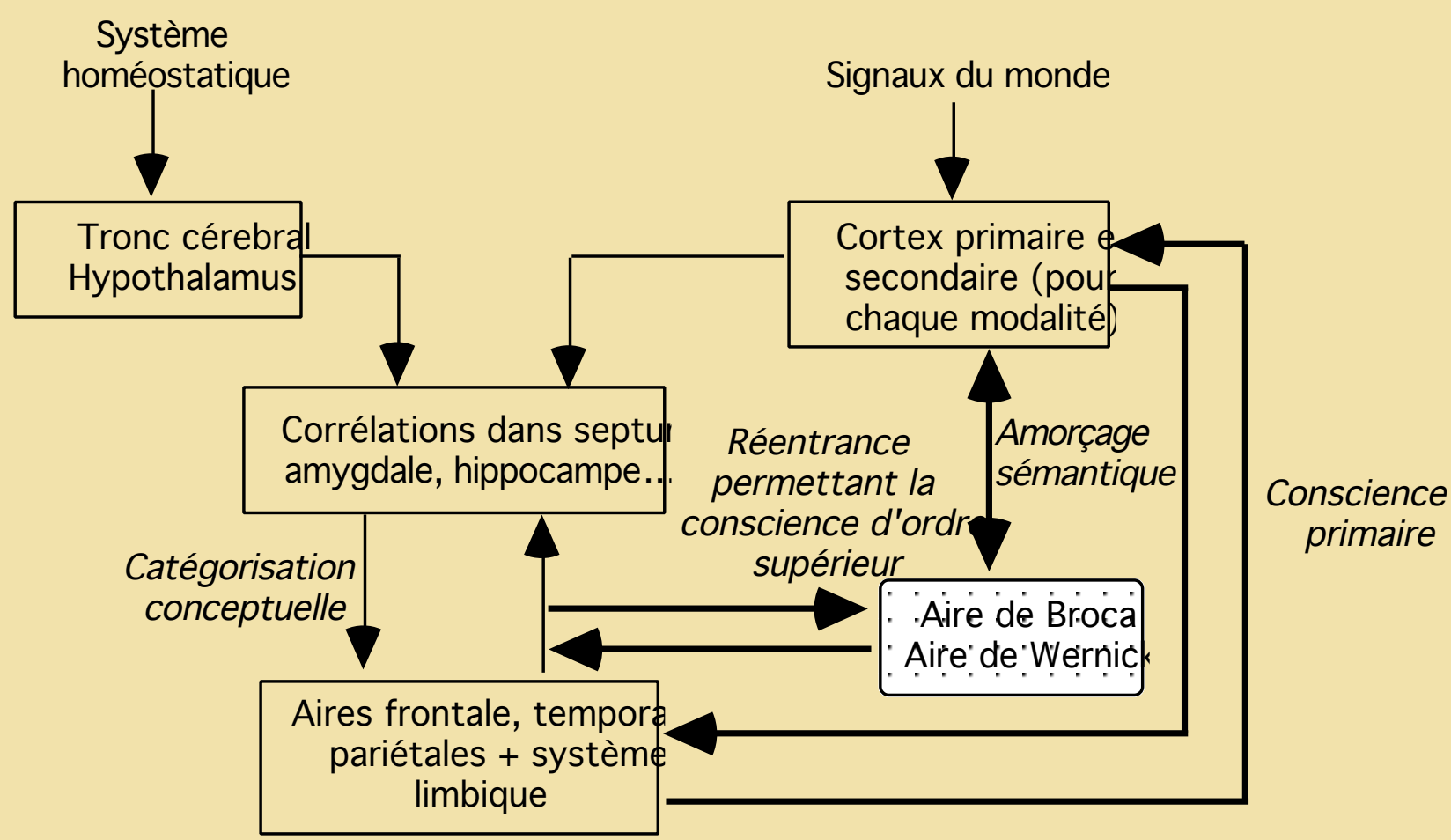
Rôles respectifs des concepts, des symboles, de la sémantique et de la pragmatique vis-à-vis du langage



Catégorisations



La conscience de haut niveau



Cardon

- « Approche constructible de la conscience artificielle » Alain Cardon, 2003, Automates intelligents, Paris
- principales hypothèses
  - *une pensée artificielle est calculable*
  - *elle nécessite l''interaction physique d'un corps matériel avec son environnement.*

Modélisation de phénomènes complexes

- Fondamentalement différente des modélisations équationnelles ou formelles
  - *définition des éléments générateurs des mouvements de base du phénomène*
  - *définition des règles de communication et de synchronisation entre ces éléments*
  - *mise en mouvement des éléments de base (évolution temporelle du phénomène)*
  - *observation de ces mouvements*

- Complexité← calculs et interactions (¬ syntaxe du modèle)

Résumé

- composants de base = agents actifs, proactifs (capables, de leur fait propre, de mouvements et de communications), symboliques, évolutifs et communicants
- regroupés en ensembles et sous-ensembles (nommés *agents aspectuels*), communicants, au fonctionnement réflexe, automatique, inévitable
- *agents morphologiques (régulateurs - méta)* représentent l'état et le fonctionnement des agents aspectuels
- Ces deux systèmes sont co-actifs et s'influencent réciproquement (par l'intermédiaire d'entités informationnelles). La stabilisation de l'ensemble correspond à un « *état de pensée* ».

Construction d'une « pensée artificielle »

- **Émotion**
  - ***non consciente, réponse comportementale automatique visant à s'adapter à la situation***
- **Sentiment**
  - ***conscient, déclenché par une émotion***
- ***conscience noyau***
  - ***perception de soi-même en état d'appréhension du monde (analogie forte avec la conscience de 1er niveau d'Edelman).***
- **La *conscience étendue* est alors la mise en situation de l'individu dans la temporalité**

Conscience et Intelligence artificielle(s)

Vues par Jacques Pitrat

Limites de l'intelligence humaine

- Lenteur des neurones (et nombre limité)
- Impossibilité de modifier la structure de son cerveau
- Mémoire de travail très limitée
- Capacité d'expliquer ce qu'on fait (réflexivité) mais incapacité de méta-expliquer (pourquoi on a des intuitions)
- Inaccessibilité de l'inconscient

Quelques idées de base

- Les chercheurs en IA ont deux défauts : *trop intelligents et pas assez paresseux*
- L'IA est le problème le plus difficile auquel l'homme s'est attaqué
- L'homme n'est peut-être pas assez intelligent pour le résoudre
- Il faut s'aider des systèmes d'IA eux-mêmes
- → amorçage

Idées importantes…

- META : Après résolution pb, généraliser et apprendre
- On ne copie pas l'intelligence humaine mais il est souvent bon de s'en inspirer
- Cognition artificielle ≠ cognition humaine
  - *Certaines capacités cognitives artificielles sont inaccessibles aux humains (et vice versa)*
- N'oublions pas l'IA forte ! *(AFIA 100e)*
  - *Importance des fonctionnalités de la conscience pour l'amorçage*

Conscience artificielle

- Conscience réflexive
  - *Apprendre (s'observer)*
  - *S'adapter (se modifier)*
- Conscience morale
  - *Autonomie*
  - *Choix*
- Certains aspects réalisés dans CAIA
  - *Système général de résolution de problèmes donnés sous forme de contraintes*
- Décrits dans Artificial Beings : The Conscience of a Conscious Machine (ISTE & Wiley, 2009)

CAIA (Chercheur Artificiel en Intelligence Artificielle)

- Déclarativité et réification
  - *Modification simple par changement de la valeur d'une variable*
- Pile des appels de fonctions
  - *Arrêts sans problèmes*
  - *Détection d'anomalies*
- Modification dynamique
  - *Création et compilation de nouveaux programmes*

Avantages des systèmes artificiels « conscients »

- Analyser ce qu'on sait
- S'observer en train de fonctionner
- Méta-combinatoire
- Immortalité

+ *éventuellement* meilleure compréhension de la conscience humaine

Analyser ce qu'on sait

- Accès à <u>toutes</u> les connaissances (≠*humains*)
- Compréhension
- Forme déclarative
  - *Analyse, création, modification plus faciles*
- Transformées sous forme procédurale
  - *Utilisation de méta-connaissances*
  - *450 000 lignes de C (10 000 règles conditionnelles)*
  - *Efficacité*

S'observer en train de fonctionner

- Différence fondamentale entre humains et machines : l'inconscient et la mémoire
  - *On connaît certaines étapes de nos raisonnements, mais on ne sait pas pourquoi on y a pensé*
  - *CAIA peut rendre tout « conscient »*
- Détection d'anomalies
- Interruptions et reprises aisées
- Trace →Explication, méta-explication

Conscience morale

- Autonomie
  - *Capacité de s'enrichir de sa propre expérience*
  - *Modèle réflexif nécessaire*
- Respect des valeurs et des principes
  - *Évolutifs chez l'homme (3 niveaux)*
  - *Fixes (et ≠) chez les machines*
- Impossibilité de prévoir tous les comportements d'un programme d'IA
- Nécessité d'une surveillance robuste
  - *Validation et contrôle ?*

Méta-combinatoire

- Chaque méthode de CAIA est associée à
  - *des déclencheurs potentiels*
  - *des conditions qui peuvent l'interdire*
  - *des priorités qui déterminent quand l'utiliser*
- → Combinatoire sur les méthodes elles-mêmes (+ explication)
- Exemple :
  - *Trouver tous les nombres m et n positifs et inférieurs à 10^{18} tels que:*
    $4*m + 3*n^2 = 817\ 401\ 078\ 957\ 542\ 034$

Immortalité

- Facilité de reproduire un système
- Copies identiques ou légèrement différentes
  - *Prise de risques*
  - *Test de divers variantes*
  - *Adaptation au problème*

Langage, conscience et Intelligence artificielle(s)

CARAMEL

Compréhension « automatique » des langues ?

Donner à un ordinateur les capacités cognitives lui permet-tant de se comporter *comme s'il comprenait (*représentation de la situation décrite*)*

a) Connaissances

Mots, structures de phrases, sens des mots et des phrases, usages *(lexiques, dictionnaires, grammaires, encyclopédies…)*

b) processus + architecture

Mises en œuvre informatiques de ces différents niveaux

Relations entre ces différents modules
(le problème le plus délicat !)

Ambiguïtés conscientes ou non

- a) Intentionnelles « il était une foi, la mienne » (Raymond Devos, *la vie d'un moine racontée par lui-même*).
- b) Non perçues consciemment (comprendre l'intention communicative !)

« Si vous voulez des enfants, adressez-vous à Monsieur le Curé » *(avis aux dames catéchistes, lu dans une église)*

C) [Siméon, le fils de Joannis, est curé] Interactions

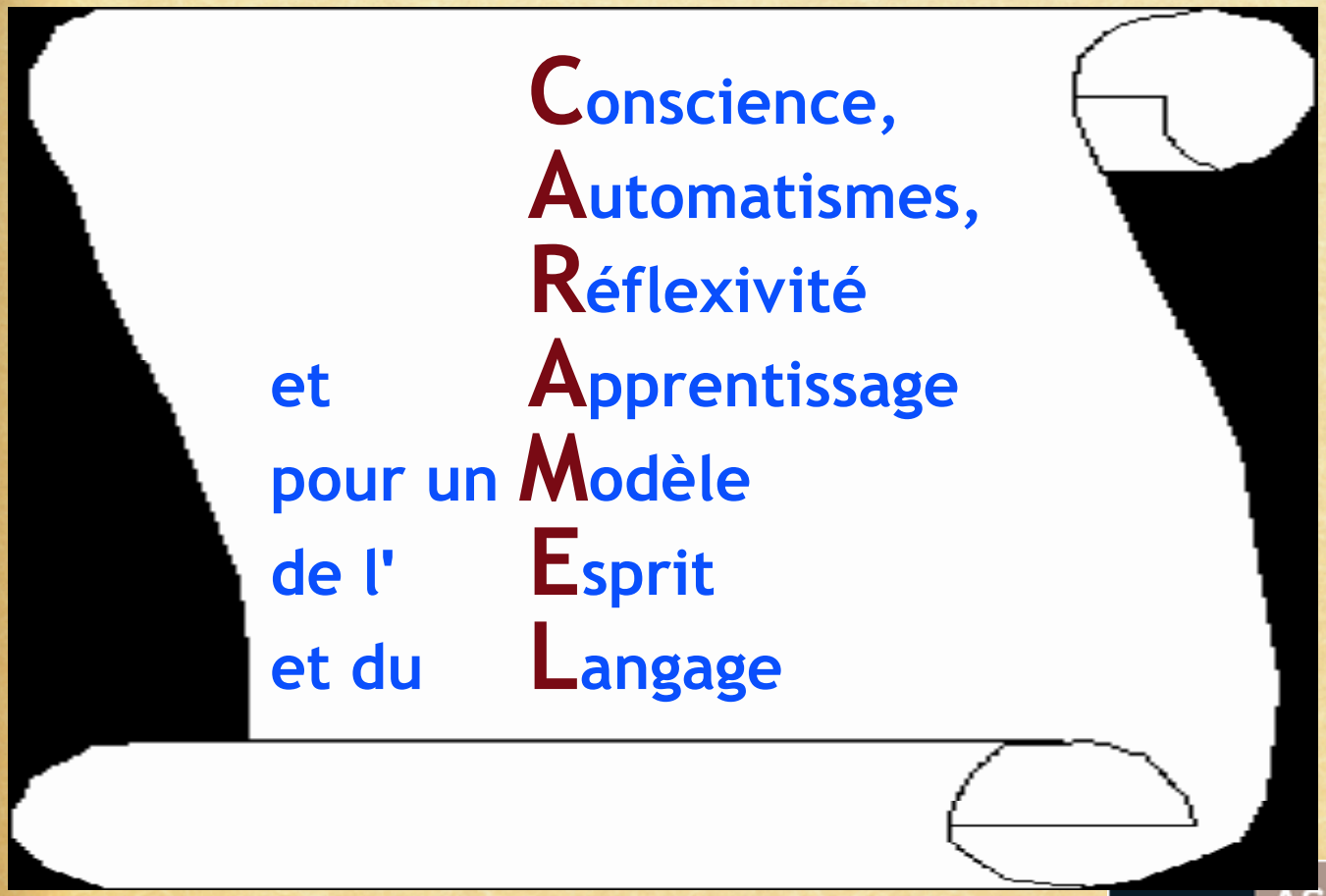
« - Mon père, dit Joanis à son fils, je suis en grand souci.

- À propos de quoi, mon fils? fit Siméon à son père. » *(Jean Anglade, Les Bons Dieux)*

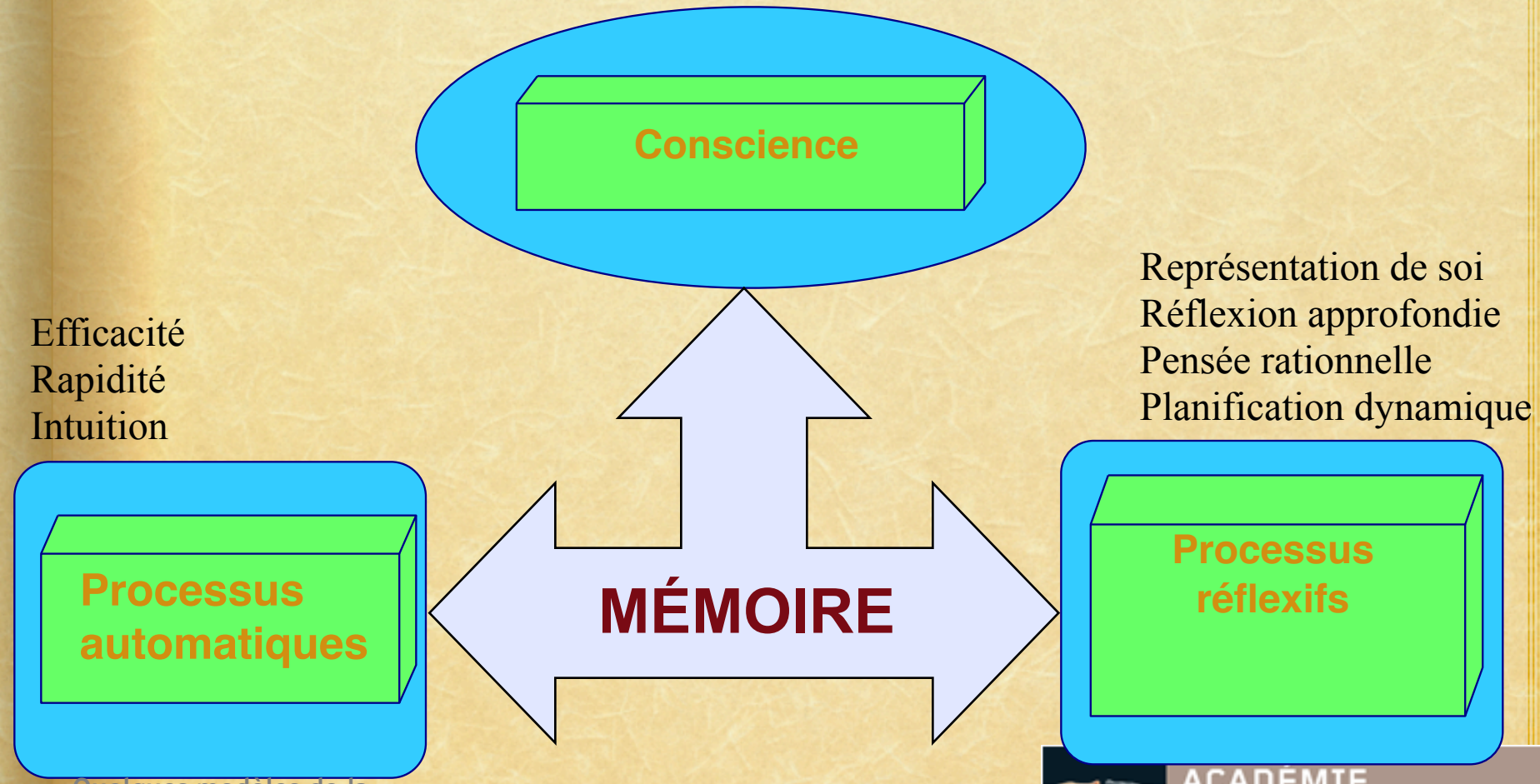
Quels processus ?

- **Phonétique, relations avec autres langues**
  *Donnez le **si**, il pousse un **if***
  *Faites le **tri**, il naît un **arbre***
  *Jouez au **bridge**, et le pont **s'ouvre**… (Boris Vian)*
- **Rôle des lois de l'arithmétique dans la compréhension**
  *En mettant les bouchées **doubles**, on fait en **6** mois ce qui devait l'être en **18** (dit par un ministre à la radio)*
- **Le texte lui-même décrit comment le comprendre**
  *« Et il m'a dit, ajouta-t-il, **en jouant de petits accords aux endroits où je mettrai des points**, que Chécoavins avait laissé. Trois enfants. Sans mère. Et que la profession de Chécoavins. Etant impopulaire. La génération montante des Chécoavins. Etait dans une situation très difficile » (Dickens, La maison d'Apre-Vent)*

CARAMEL : un modèle informatique



Principaux composants de CARAMEL



Caractéristiques du carnet d'esquisses

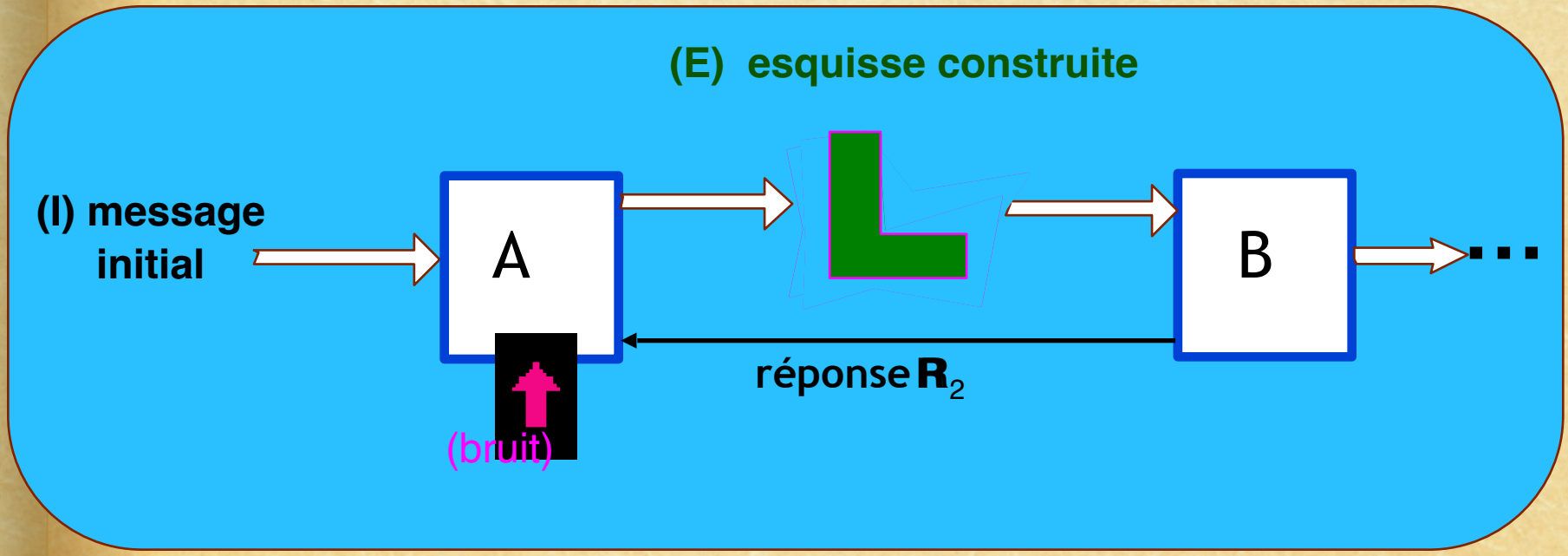
- Une extension des tableaux noirs pour les processus « inconscients »
- Relations entre processus
  - *tient compte des rétroactions des niveaux supérieurs*
  - *conserve l'encapsulation des processus (rétroaction possible entre processus qui ne se connaissent pas)*
- Esquisses construites dans le *carnet d'esquisses* et améliorées continûment

Rétroactions

- Les processus sont considérés sous deux points de vue

  ils produisent un résultat donné (une esquisse)

  ils retournent une réponse *numérique* qui indique leur degré de confiance envers leur propre résultat



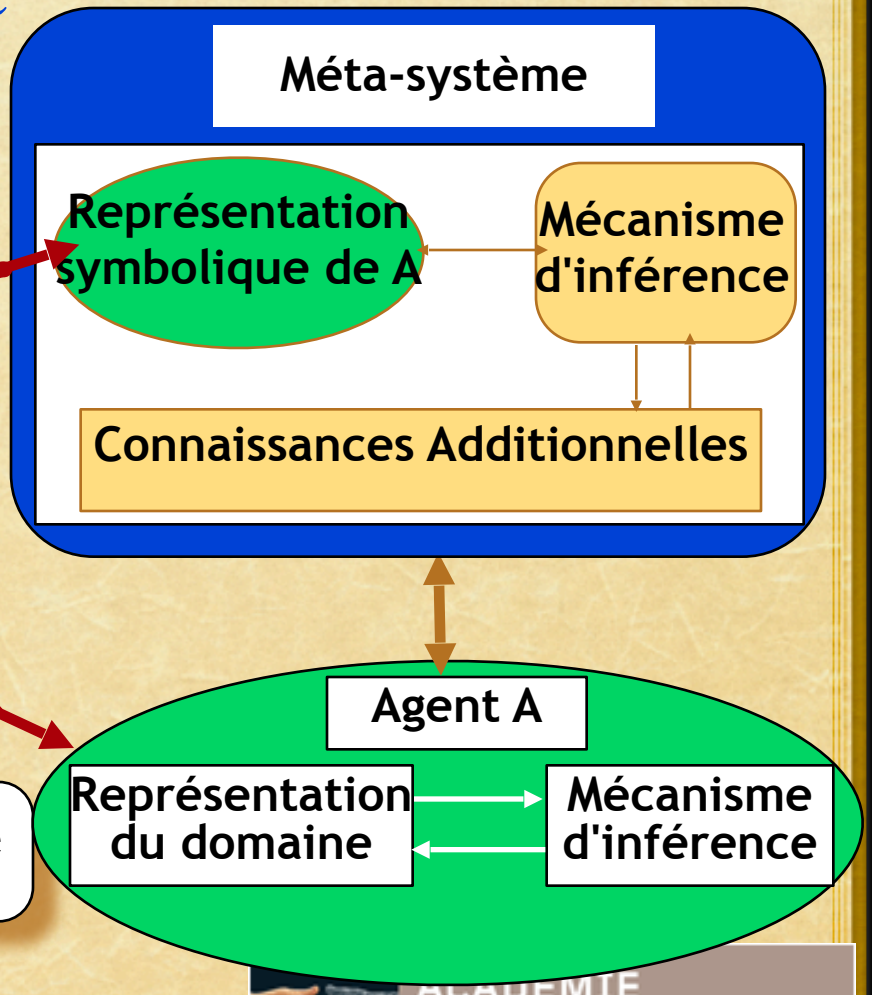
Critères / compréhension

- Résultats stables dans le carnet d'esquisses — et *qui le méritent* — écrits dans la *mémoire à court terme*
  - *sentiment de compréhension (stabilité du carnet d'esquisses)*
  - *sentiment d'ambiguïté : une « instabilité durable » (oscillations entre configurations stables)*
  - *sentiment de contradiction (résultat opposé à une attente ou une information consciente)*
  - *sentiment d'échec (un sous-but ne peut être atteint ⇒ problème conscient ⇒ planification explicite)*

Systèmes Réflexifs

- Un méta-système utilise un modèle symbolique de A pour raisonner à propos de A
- Connexion causale entre le système et sa représentation

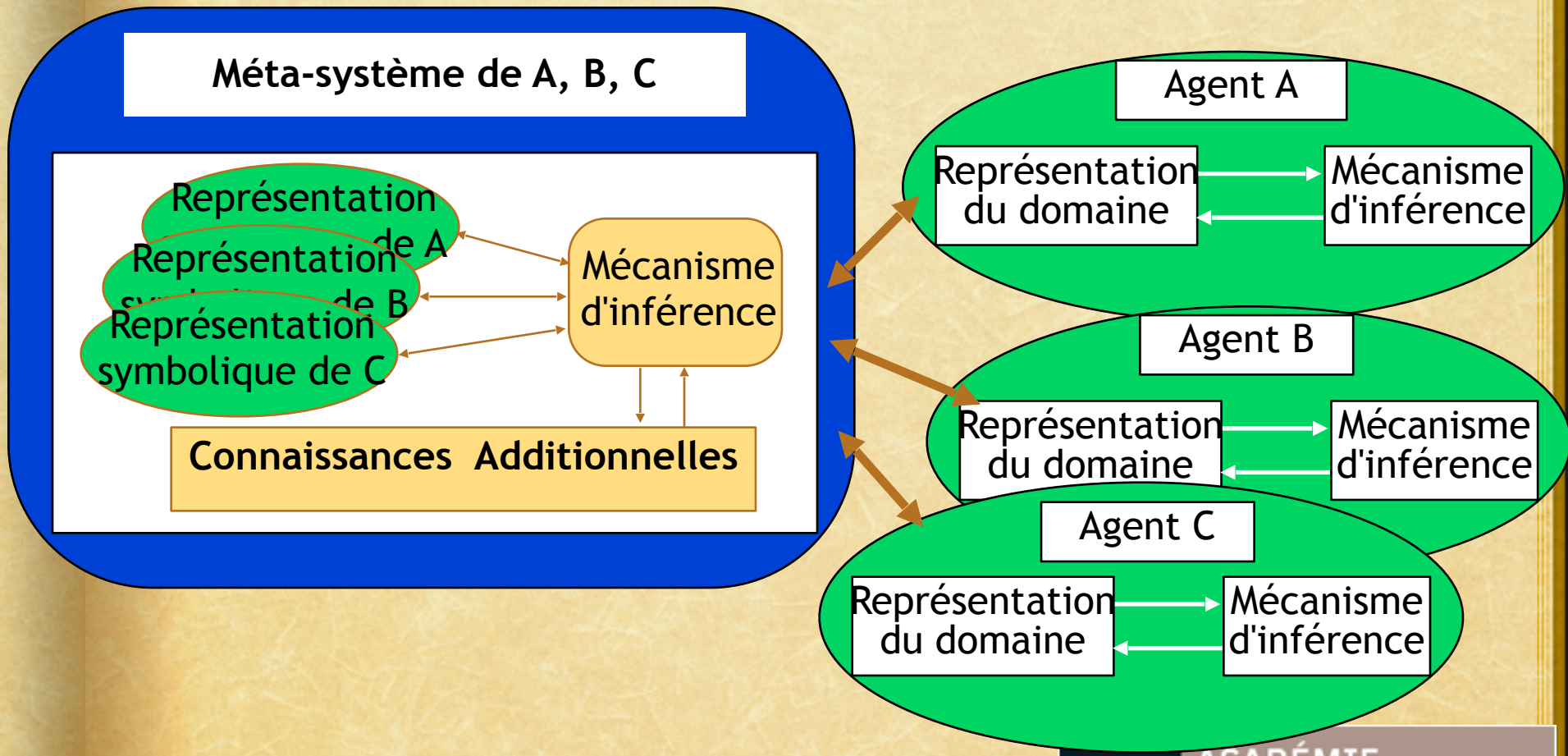
  (la représentation doit être toujours *fidèle*)
- La méta-représentation est sélective et spécialisée



La méta-représentation est partielle

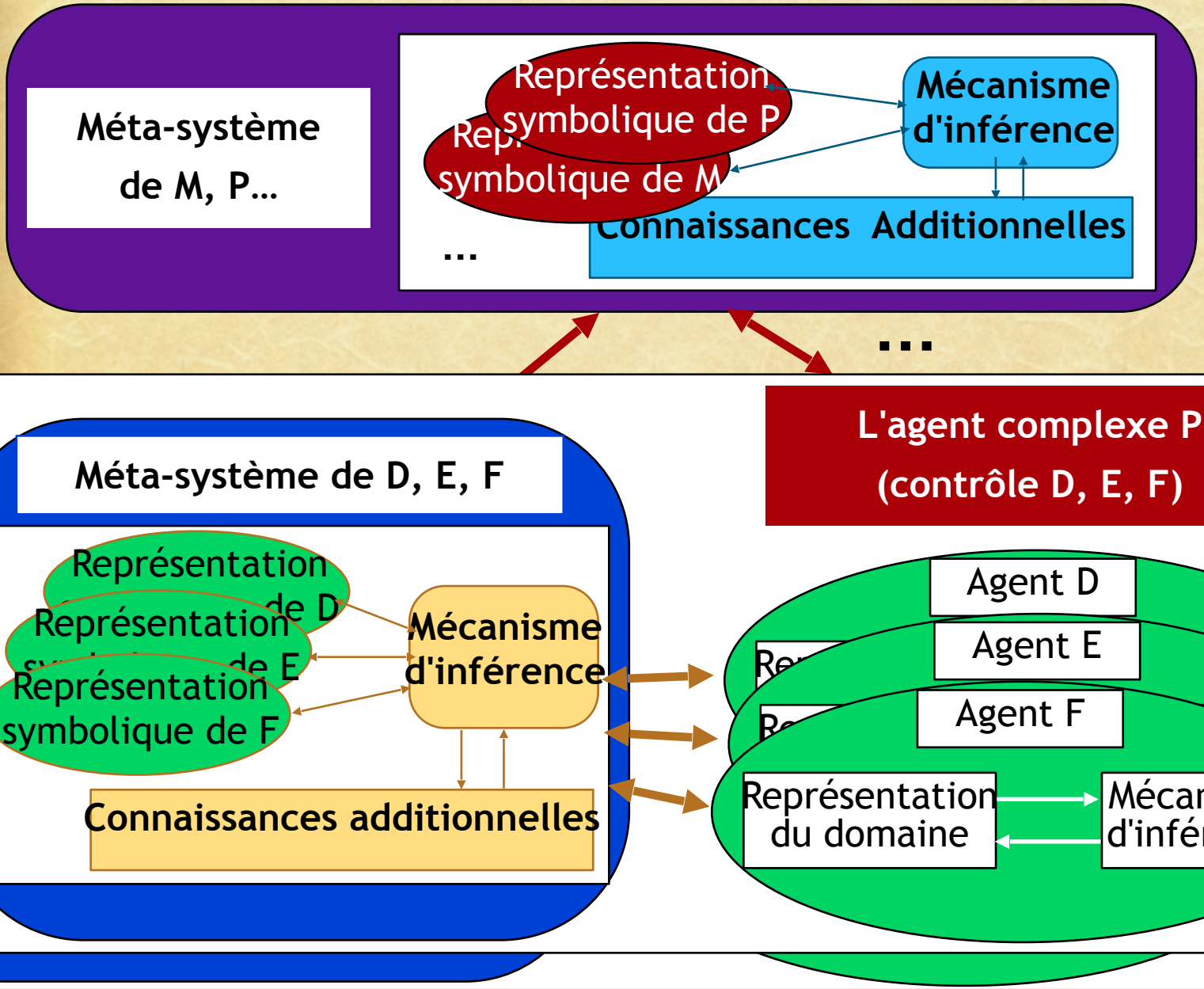
Méta-systèmes récursifs

un méta-système contrôlant plusieurs agents



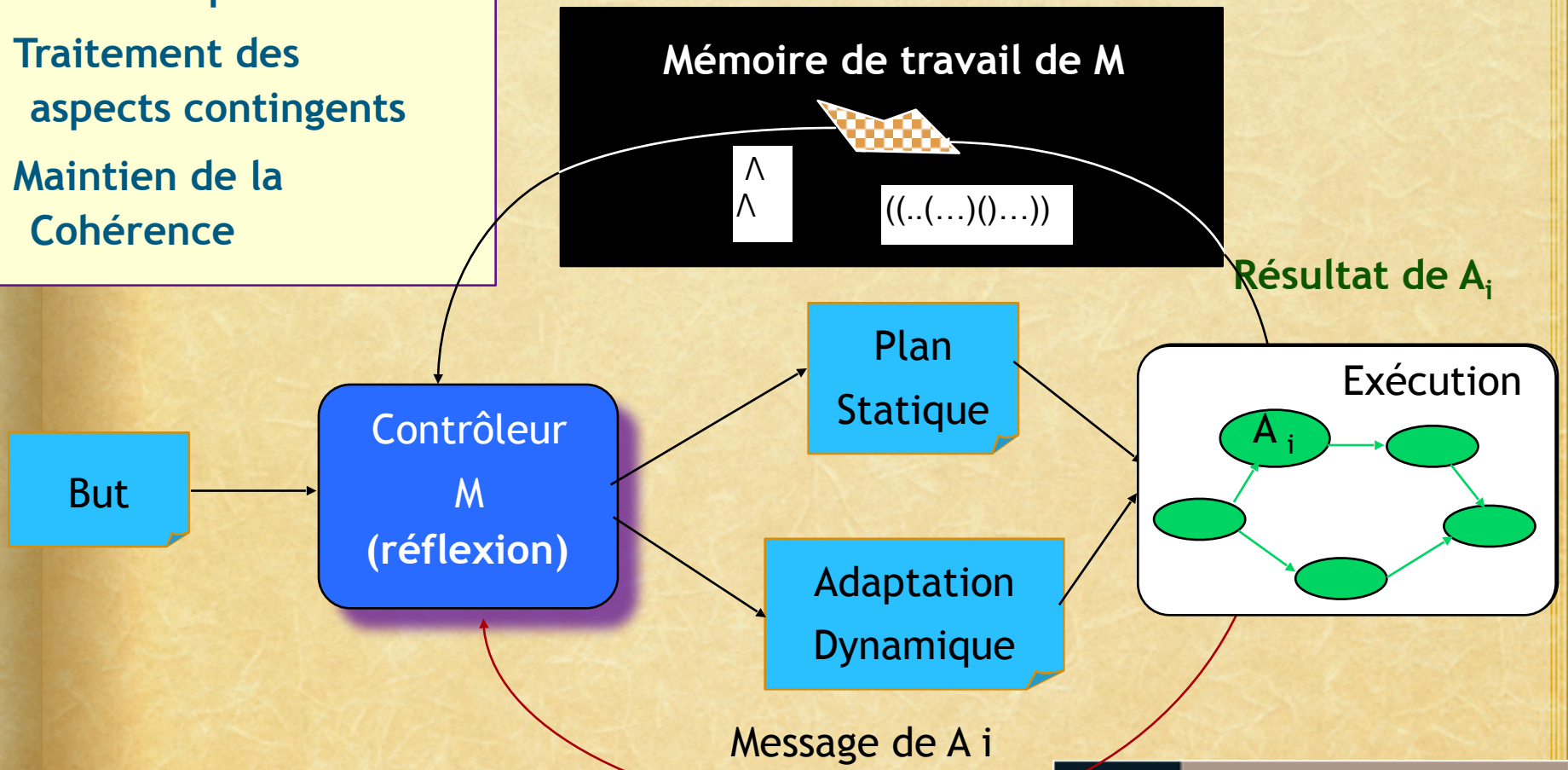
Méta-systèmes récursifs

un méta-système : un agent usuel

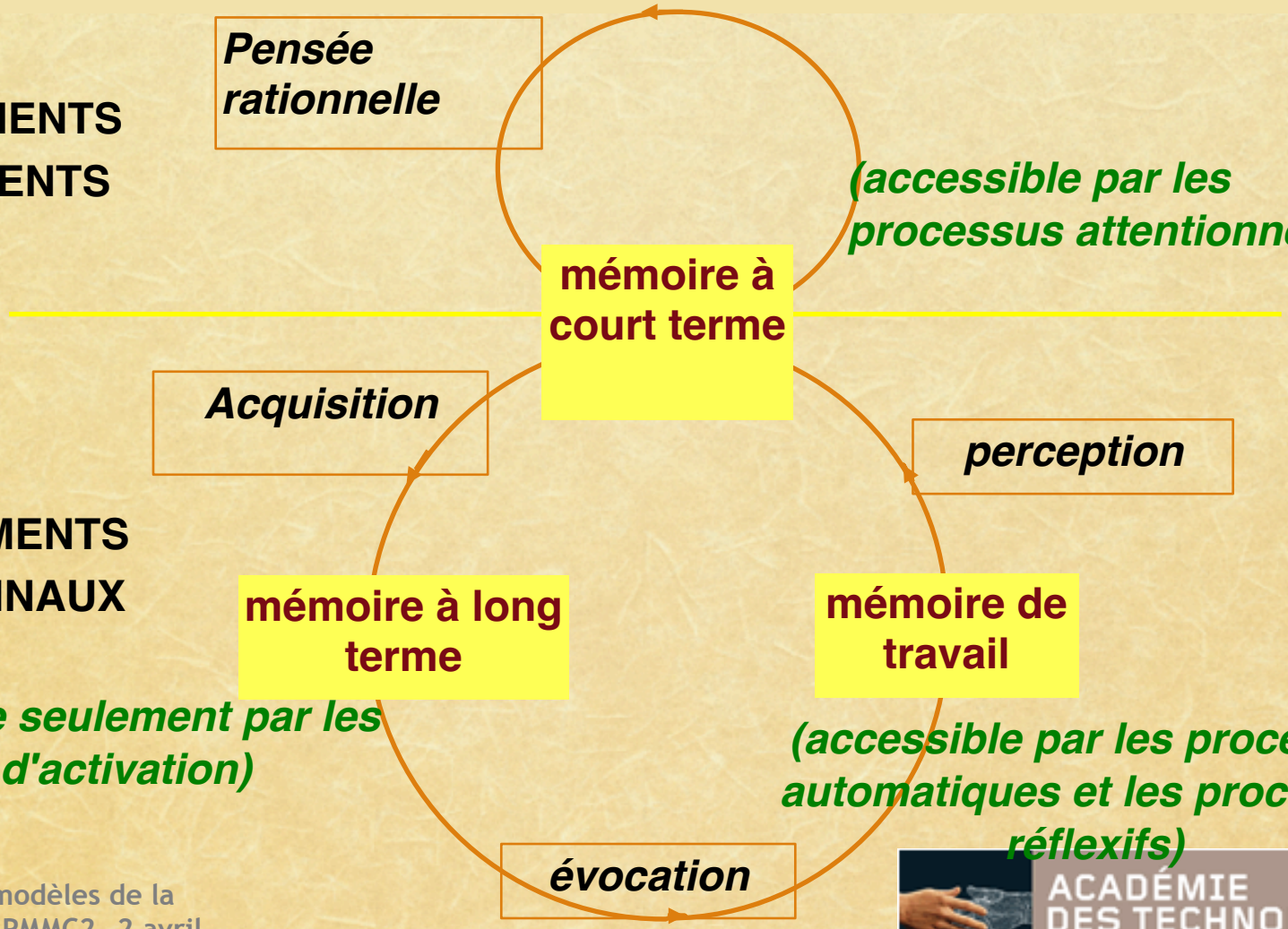


Le raisonnement réflexif de CARAMEL

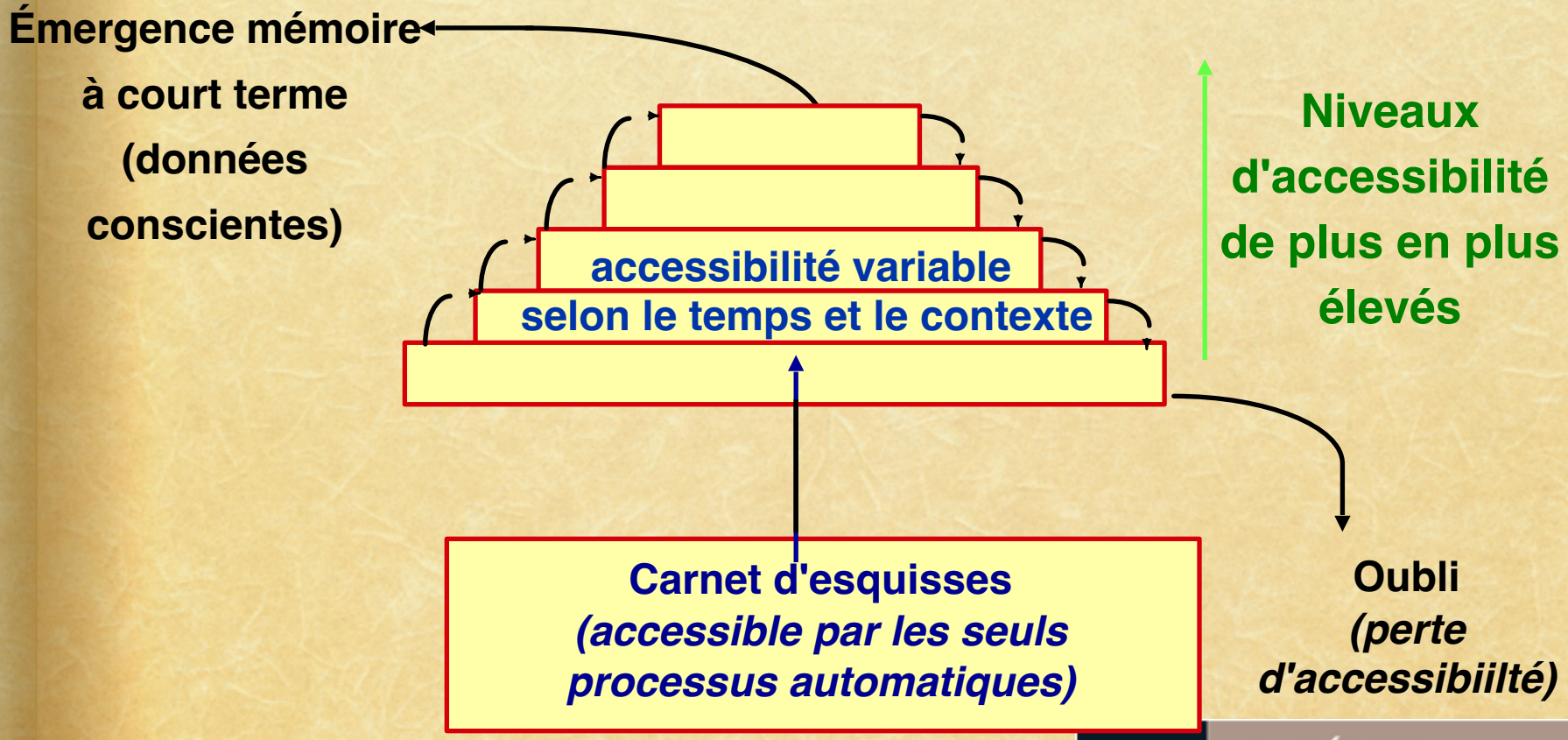
- Intégration de résultats partiels
- Traitement des aspects contingents
- Maintien de la Cohérence



Modèle de mémoires



Mémoire de travail



Mémoire à long terme

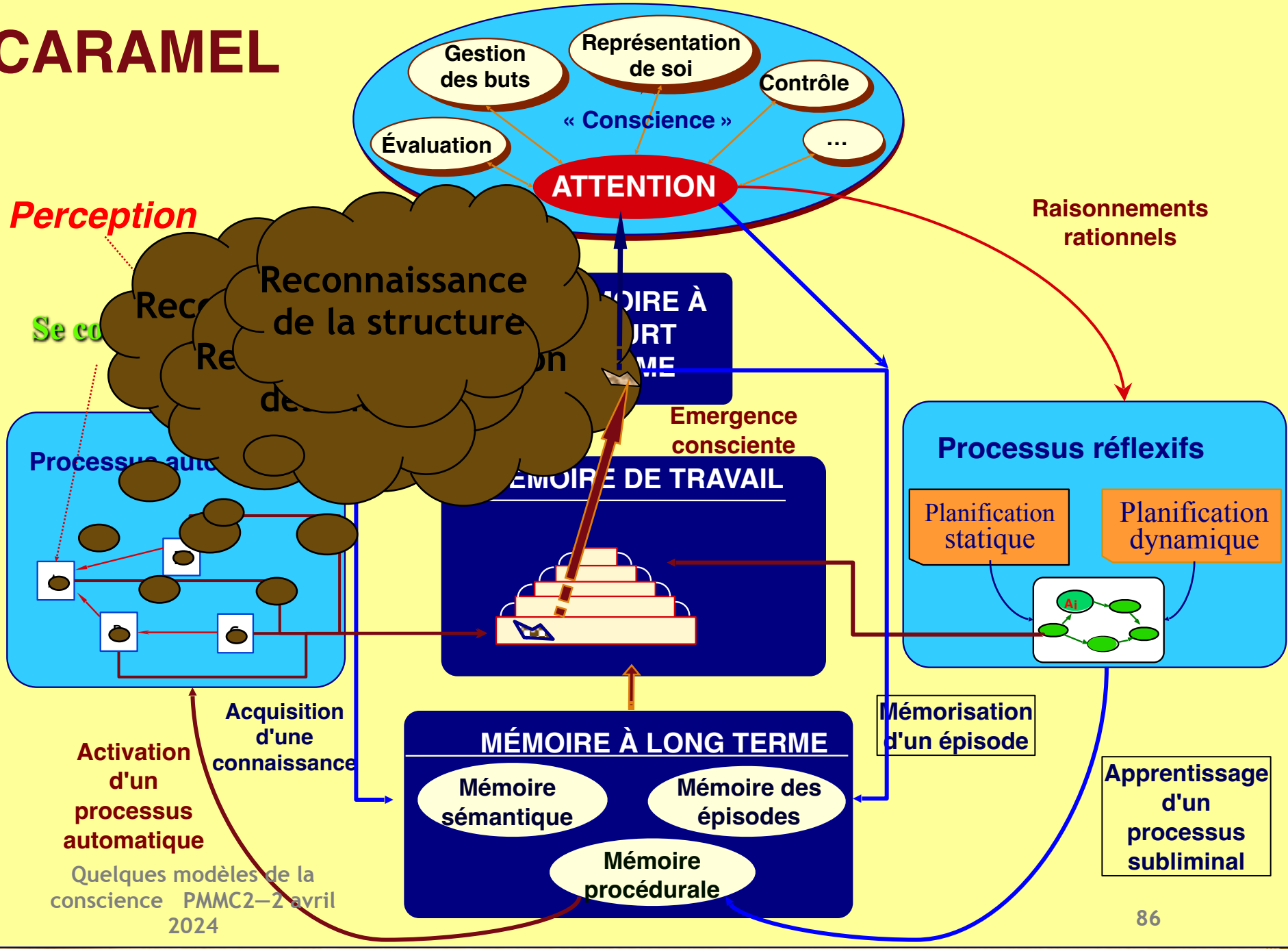
Divisée en trois mémoires différentes

Mémoire sémantique
(représentation de concepts, propriétés et relations : graphes conceptuels)

Mémoire des épisodes
(représentation, stockage et généralisation d'événements préalables)

Mémoire procédurale
(représentation et stockage des processus automatiques)

CARAMEL


Gestion des buts
Représentation de soi
Contrôle
« Conscience »
Évaluation
…
ATTENTION
Perception
Reconnaissance de la structure
Raisonnements rationnels
Emergence consciente
Processus réflexifs
Planification statique
Planification dynamique
Ai
MÉMOIRE DE TRAVAIL
MÉMOIRE À LONG TERME
Mémoire sémantique
Mémoire des épisodes
Mémoire procédurale
Acquisition d'une connaissance
Activation d'un processus automatique
Mémorisation d'un épisode
Apprentissage d'un processus subliminal

Conclusion

- Énorme potentiel de l'IA
  - *Implémentation et expérimentation nécessaires*
- Deux problèmes pour l'IA forte
  - *Intelligence humaine*
  - *Structure de la recherche*
    - Temps consacré
    - Long terme
- Espoirs : sciences cognitives et amorçage
  mais, *une perspective à 100 ans…*

Besoins pour une IA forte

- Prendre du recul
- Changer dynamiquement de mode de raisonnement
- Joindre plusieurs sources d'informations
- Trouver des analogies entre univers différents
- Avoir la capacité de développer des théories puis de les vérifier par l'expérimentation.
- Avoir ***envie*** de faire quelque chose

Problèmes ouverts

- Monde ouvert vs. Monde fermé
- Veille difficile : un millier de start-up en IA aux EU (seulement une soixantaine en France)
- Vision stratégique à long terme des GAFA
- Nouveau rôle de l'expert : poser les bonnes questions (et non trouver les bonnes réponses)

ACADÉMIE DES TECHNOLOGIES

POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

« IA forte » science fiction ou question de calendrier ?

- Pas d'impossibilité prouvée
- L'IA n'est pas une technologie comme les autres (*Autonomie*, *apprentissage*, *auto-modification*)
- Les risques du transhumanisme
- Problème de la validation :
  Comment garantir les propriétés d'un tel système ?
  Comment conserver le contrôle ?

Voir aussi

- Groupe de travail de l'Académie des technologies « vers une technologie de la conscience ? »
- Présidents : Gérard Sabah, Philippe Coiffet
- Une quinzaine de participants de l'Académie + invités extérieurs

ACADÉMIE DES TECHNOLOGIES

POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

Documents disponibles

Accessibles à l'adresse : gscns.free.fr/

- Rapport du groupe : *RapportGT-conscience.docx*
- Présentations effectuées lors des réunions (membres du groupe et experts invités)
- Comptes rendus de toutes les réunions
- Un glossaire (78 termes)
- Une bibliographie (221 références)
- Divers autres documents pertinents, en particulier *robot de compagnie*, *robot militaire*

Merci !

DISCUSSION…