Confiance dans les systèmes utilisant l'intelligence artificielle pour le contrôle et commandement des opérations aériennes C2

Bryan Fruchart

université de BORDEAUX

THESE PRESENTEE
POUR OBTENIR LE GRADE DE

DOCTEUR
DE L'UNIVERSITÉ DE BORDEAUX

Ecole Doctorale Sciences Physiques et de l'Ingénieur
Automatique, Productique, Signal et Image, Ingénierie Cognitique

Par Bryan Fruchart

Confiance dans les systèmes utilisant l'intelligence artificielle pour le contrôle et commandement des opérations aériennes C2.

Equipe de recherche Cognitique et Ingénierie Humaine (CIH) - IMS
Sous la direction du Professeur Benoît Le Blanc

Soutenance orale le 16 décembre 2022

Membres du jury

M. CHAUDRON Laurent	DR, Theorik-Lab / Crea Ecole de l'air et de l'espace	Rapporteur
M. KHAMASSI Mehdi	DR, Institut des systèmes intelligents et de robotique	Rapporteur
Mme. SEMAL Catherine	PR. Institut polytechnique de Bordeaux	Examinateur
M. CHOPIN Philippe	Ingénieur. Thales Land & Air Systems	Examinateur
M. MARION Damien	Dr, Thales Land & Air Systems	Examinateur
M. LE BLANC Benoît	PR. Institut polytechnique de Bordeaux	Directeur

« L'intuition [...] est la source de toute vérité et le fondement de toute science.
Il faut en excepter seulement la logique, qui est fondée sur la connaissance non intuitive,
quoique immédiate, qu'acquiert la raison de ses propres lois. »

Arthur Schopenhauer

« Les conjectures sont d'une grande importance
car elles suggèrent des pistes de recherches utiles »

Alan Turing

Résumé

Nous proposons une architecture cognitive, nommée I.P.S.E.L. (Information Processing System with Emerging Logics), qui décrit le fonctionnement coopératif de plusieurs systèmes de traitement de l'information dans la perspective de mettre en œuvre une intelligence tendant à s'approcher de celle de l'Homme.

L'architecture a été construite via l'assemblage de différentes théories de sciences cognitives en un seul modèle général. Le modèle est présenté dans le formalisme des systèmes de traitement de l'information et de la cybernétique. Sa conception est le fruit d'une recherche transdisciplinaire et entre dans le spectre des travaux en architecture cognitive et en intelligence générale artificielle, deux thématiques entrant dans le champ de l'IA.

En utilisant ce modèle nous proposons une classification des systèmes d'IA. Cette classification est destinée au monde de l'entreprise et a vocation à éclaircir les différentes natures de systèmes d'IA ainsi que leurs comportements génériques. Nous proposons aussi une catégorisation des dispositions cognitives d'un utilisateur de système d'IA. Ces deux ensembles nous permettent de discriminer des situations d'interactions particulières entre un utilisateur et un système utilisant l'IA. Pour chacune de ces situations nous discutons alors des conditions qui permettent de mitiger les risques d'une utilisation inappropriée.

Le modèle ainsi proposé a aussi vocation à servir de cadre théorique pour l'analyse des comportements artificiels. Nous montrons son utilité dans ce sens en nous en servant pour enrichir les discussions qui portent sur : les capacités des machines dotées d'intelligence artificielle, l'intelligence artificielle de confiance et l'intelligence générale artificielle.

Pour décrire le modèle sous une forme tangible, nous proposons le pseudo-code d'un programme informatique qui montre l'emboîtement des différents modules de l'architecture IPSEL.

La perspective qui consiste à concevoir un système de traitement de l'information autour d'un principe de logique émergente semble prometteuse pour concevoir un véritable agent artificiel, autonome et intelligent. Notre modélisation peut donc être considérée comme une proposition du fonctionnement d'une AGI (Artificial General Intelligence).

Mots clés : architecture cognitive, intelligence artificielle, facteurs humains, confiance.

Abstract

We propose a cognitive architecture, named I.P.S.E.L. (Information Processing System with Emerging Logics), which describes the cooperative functioning of several information processing systems with the objective to implement an intelligence tending to approach that of humans.

The architecture was constructed by bringing together different cognitive science theories into a single general model. The model is presented in the formalism of information processing systems and cybernetics. Its design is the result of transdisciplinary research and falls within the spectrum of cognitive architecture and artificial general intelligence works, which are sub-disciplines of AI.

By using the model, we propose a classification of AI systems. This classification is intended for the business world and is intended to clarify the different natures of AI systems as well as their generic behaviors. We also propose a categorization of the cognitive dispositions of an AI system user. These two sets allow us to discriminate interaction situations between a user and a system using AI. For each of these situations, we then discuss the conditions that make it possible to mitigate the risks of inappropriate use.

The model thus proposed is also intended to serve as a theoretical framework for the analysis of artificial behavior. We show its usefulness in this sense by using it to enrich discussions about: the capabilities of machines with artificial intelligence, trusted artificial intelligence and artificial general intelligence.

To describe the model in another form, we also propose the pseudo-code of a computer system that would implement the IPSEL model as the architecture of its global processing.

The prospect of structuring an information processing system around a principle of emergent logic seems promising for designing a real artificial, autonomous, and intelligent agent that resembles humans. Our modeling can therefore also be considered as a proposal for the functioning of an AGI (Artificial General Intelligence).

Keywords: cognitive architecture, artificial intelligence, human factors, trust.

Remerciements

Je tiens à exprimer ma plus sincère gratitude envers mon directeur de thèse, Benoît Le Blanc, pour son soutien constant, ses précieux conseils et son encadrement tout au long de mon travail de recherche. Je suis également reconnaissant envers les membres du jury, pour leur temps et leur expertise consacrés à l'évaluation de ma thèse.

Je remercie la présidente du jury Catherine Semal pour sa confiance, ses encouragements et conseils ainsi que pour la bienveillance et sollicitude dont elle a fait preuve à mon égard.

Je remercie Laurent Chaudron pour son implication et ses nombreux retours et pour les connaissances complémentaires qu'il m'a apporté lors de nos discussions.

Je remercie Mehdi Khamassi pour le temps engagé dans l'évaluation de mon travail ainsi que pour les compléments éclairés qu'il a pu me fournir dans son rapport et pendant la soutenance orale.

Je remercie Damien Marion, qui avant d'être examinateur de mon jury a été un collègue de laboratoire avec lequel nous avons eu nombre d'échanges enrichissants et stimulants.

Je remercie Philippe Chopin pour sa confiance en tant que représentant de la société Thales et pour les heures passées au téléphone à discuter des systèmes de demain.

Je remercie l'ensemble du personnel de l'ENSC pour son accueil et les moyens mis à disposition.

Je remercie également la société Thales Land & Air System, notamment Thierry Bon, à l'initiative de ce travail doctoral.

Enfin je remercie chaleureusement mes proches et parents pour leur soutien moral et leur patience.

Table des matières

Liste des figures

Organisation du manuscrit

Le premier chapitre (Chapitre 1 : Introduction générale) est une introduction dans laquelle sont présentés le sujet de recherche et le contexte industrialo-technologique dans lequel l'étude s'inscrit.

Le second chapitre (Chapitre 2 : État de l'art) rapporte les différentes études bibliographiques qui ont été menées durant notre travail. Les sujets de ces études concernent la confiance Homme-Automation (Partie 2.1), le domaine de l'intelligence artificielle (Partie 2.2), la cognition humaine (Partie 2.3) et l'architecture cognitive (Partie 2.4).

Nous décrivons ensuite le modèle que nous proposons dans le troisième chapitre (Chapitre 3 : Modélisation I.P.S.E.L.). D'abord en présentant notre positionnement (Partie 3.1), puis en décrivant l'architecture fonctionnelle (Partie 3.2). Nous présentons également une description subjective de la mise en marche d'une machine dotée d'une telle architecture (Partie 3.3) et conjecturons sur des principes et propriétés attribuables aux systèmes dont la cognition suit la modélisation IPSEL (Partie 3.4).

En quatrième chapitre (Chapitre 4 : Utilisation du modèle), nous utilisons le modèle que nous avons construit pour catégoriser les types de comportements cognitifs naturels et artificiels (Partie 4.2), pour analyser les interactions Homme-IA (Partie 4.3) et discuter la notion d'intelligence artificielle de confiance (Partie 4.4).

Pour le cinquième chapitre (Chapitre 5 : Vers une intelligence générale artificielle), nous présentons un pseudo-code montrant l'organisation informatique de l'architecture que nous avons développée ainsi que des pistes d'implémentations technologiques (Partie 5.1). Nous finissons ce chapitre par une discussion sur la notion d'intelligence générale artificielle (Partie 5.2).

Chapitre 1 : Introduction générale

Une très récente étude, datée de janvier 2022, estime le marché de l'intelligence artificielle (IA) à 342 milliards de dollars et prédit 500 milliards pour 2024 [Tazrout 2022]. Selon le centre d'étude des risques existentiels de l'université de Cambridge, l'IA expose les sociétés à de nombreuses menaces [Brundage et al. 2018], voire pire, elle initierait le déclin de l'humanité selon le brillant physicien Stephen Hawking [Cellan-Jones 2014].

Lorsque notre travail de recherche commence en 2018, le domaine de l'intelligence artificielle (IA) est en pleine effervescence. En ce début du XXI siècle, des techniques algorithmiques théorisées dans les années 1950 deviennent réellement efficaces. Les applications du *machine learning* et des réseaux de neurones artificiels se démocratisent, notamment grâce aux volumes de données qu'Internet expose et grâce à la puissance de calcul des ordinateurs qui s'est considérablement accrue.

Les nouvelles technologies de l'ère de l'information ont le potentiel de changer la face du monde tout comme l'écriture, l'imprimerie et l'électricité l'ont fait auparavant. Nul ne saurait le réfuter pour l'informatique et Internet qui conduisent aux télécommunications modernes. Pour l'IA, s'il était encore possible d'en débattre en 2018 les récentes applications qui pullulent dans la plupart des secteurs industriels en 2022 finiront de convaincre les plus sceptiques [Lee & Yoon 2021][Pandl et al. 2020].

Au-delà des résultats académiques et de l'intelligence que les systèmes de transistor ont démontré en battant l'Homme dans la plupart des jeux numériques, (jeu d'échec [Maharaj et al. 2022], jeu de go [Silver et al. 2017], poker [Zhang et al. 2022], starcraft 2 [Tual 2019]), l'intelligence artificielle permet désormais de nombreuses applications concrètes qui appartenaient à la science-fiction il y a encore quelques décennies.

Les machines ont acquis des capacités équivalentes à celles de leurs opérateurs de chair et d'os. Elles peuvent reconnaître, identifier et classer des formes, des objets, des visages et tout autre motif spatio-temporel dans un ensemble de données quel qu'il soit. Elles peuvent contrôler avec précision un bras robotisé, une voiture, un drone, un avion ou tout autre appareillage électronique. Les machines peuvent également raisonner sur des problèmes complexes comme diagnostiquer une pathologie à partir d'un ensemble de symptômes, trouver les formes de repliement de molécules, ou toute autre application nécessitant la

manipulation et le recoupement de connaissances relationnelles. Leurs derniers accomplissements les plus impressionnants sont certainement ceux qui ont trait au langage. Les systèmes artificiels peuvent traduire à peu près toutes les principales langues parlées dans le monde. Ils peuvent synthétiser des articles et des livres, obéir à des consignes formulées dans le langage courant. Ils peuvent produire des voix et des discours artificiels. Ils peuvent générer des images de n'importe quel objet ou combinaison originale d'objets à partir d'une instruction exprimée en langue humaine. De manière générale, ils peuvent répondre à des questions complexes, ce qui, pour beaucoup, sera facilement assimilé à de la compréhension.

L'intelligence artificielle a vocation à imiter et à simuler l'intelligence humaine. Tout comme son modèle naturel, l'intelligence artificielle est imparfaite. Aucune intelligence n'est absolue car les erreurs permettent d'apprendre et l'omniscience n'existe pas. Les jugements des machines sont parfois biaisés par leurs expériences inévitablement restreintes. Les raisonnements des machines sont parfois erronés à cause de leurs connaissances inévitablement bornées. Il ne semble pas raisonnable d'en attendre mieux, tout système, qu'il soit naturel ou artificiel, ne pourra disposer que d'une intelligence limitée par ce qui lui est permis de percevoir, de représenter et de modéliser.

Le déploiement dans nos sociétés de systèmes artificiellement intelligents et ce constat quant à l'imperfection de leur intelligence amène tout un chacun à s'interroger. Des questionnements du type « Devons-nous continuer à développer ces technologies ? » ont initié des discussions éthiques, philosophiques et futurologistes. D'autres interrogations comme « Comment bien utiliser ces technologies ? » ont initié des processus de réglementation, de normalisation et de législation. Enfin, des questions du type « Quand faire confiance à ces technologies ? » ont orienté les efforts de recherche vers la conception de systèmes intelligibles, explicables, ou capables de justifier leurs productions.

Notre étude, réalisée pour le compte de la société THALES Land & Air System, au laboratoire IMS de l'Institut Polytechnique de Bordeaux, s'inscrit dans ce contexte. Celui du développement d'un nouvel outil technologique, de son déploiement dans nos groupes sociaux et de ses premiers usages à grande échelle, notamment dans des secteurs d'activités critiques comme celui des opérations militaires.

Le sujet de ce travail doctoral est le suivant : « **compréhension de situation complexe, confiance dans les systèmes utilisant l'intelligence artificielle** ». Le cadre applicatif visé

est celui du commandement et contrôle des opérations aériennes et l'intégration des technologies de l'IA dans ses processus, avec un horizon placé en 2035. Le commandement des opérations aériennes est une activité qui implique de nombreux utilisateurs humains se trouvant dans des contextes hétérogènes. Un opérateur en salle d'opération, un décisionnaire à l'état-major, un pilote en phase de combat, un co-pilote en phase de navigation, tous interagissent avec des systèmes utilisant l'intelligence artificielle. Selon leur profil et la situation qui est la leur, ces différents utilisateurs ne possèdent pas les mêmes connaissances, les mêmes expériences, les mêmes contraintes ou dispositions cognitives au moment de décider d'utiliser, ou non, les systèmes d'IA censés les aider à accomplir leur objectif. De l'autre côté de l'interaction, les systèmes utilisant l'IA ont de multiples formes, sont issus de conceptions diverses, ils utilisent différentes techniques algorithmiques pour réaliser différentes fonctions. Pour les applications les plus complexes, ce sont même des systèmes de systèmes dont il est question.

La gestion moderne des opérations militaires utilise des technologies de l'information et de la communication pour améliorer le commandement et le contrôle (C2). Le C2 est une méthode et un procédé associé à une activité complexe qui nécessite un contrôle strict. Il est défini par un "*Operational Design*" qui doit conduire à l'effet final recherché. La décision peut être partagée jusqu'aux plus bas échelons si nécessaire dans une structure privilégiant l'agilité et la délégation. Les principes du C2 sont issus de l'organisation militaire, mais on note également son influence dans d'autres domaines d'application du domaine civil. Au tournant du siècle, la technologie a évolué encore plus en C4ISR, ou maintenant en C4ISR-TAR, pour Command, Control, Communications, Computers, Intelligence, Surveillance, Reconnaissance, et Target Acquisition and Reconnaissance spécifique (acquisition de cibles). Se référer à [Claverie & Desclaux 2015][Claverie & Desclaux 2016].

Le domaine stratégique du commandement et contrôle des opérations repose à la fois sur la connaissance de l'information la plus complète possible, la confiance dans les meilleures technologies qui la délivrent, et la capacité de prise de décision du commandant qui se fie à une organisation solide et efficace. Dans le contexte d'une information massive, ces trois dimensions nécessitent le développement de logiciels dits "intelligents" capables de sélectionner, fusionner et représenter des informations pertinentes et de fournir des solutions de prise de décision à haute vitesse. Ces agents sont élaborés par des grands industriels, ils progressent régulièrement vers une plus grande autonomie. Malgré ce progrès et face à une complexité croissante des situations critiques, le projet de systèmes purement

autonomes s'éloigne des perspectives réalistes à court et à moyen terme. Les experts du C2 et ces systèmes artificiels doivent de plus en plus travailler de manière collaborative, chacun apportant le meilleur de leurs compétences au duo humain-système. La notion de confiance est donc centrale pour l'I2HM (Interaction Intelligente Homme-Machine) et la collaboration entre les humains et les machines. Se référer à [Desclaux 2022].

Parmi les enjeux de l'I2HM, on note que le processus de prise de décision mis en place par les humains est radicalement différent de celui des machines intelligentes. Des architectures cognitives identiques pourraient faciliter leur communication, mais contrairement aux humains, les machines sont restreintes à des objectifs et des priorités bien définis, sans capacité d'improvisation ou d'adaptation interprétative, et sans véritable inventivité au-delà de la proposition algorithmique de solutions inattendues. Les humains, d'autre part, peuvent développer ces qualités, mais restent médiocres dans la description précise de leurs intentions, objectifs et priorités comme les machines intelligentes les exigent. De même, leurs capacités d'attention, de mémoire ou de raisonnement fiable sont fragiles et fréquemment compromises, tandis que les systèmes artificiels sont plus performants sur ces points.

De l'avis général, les machines intelligentes restent et resteront, du moins dans un avenir proche, partiellement incompréhensibles pour les humains. Il en est de même pour les humains en ce qui concerne les machines. Établir la confiance entre les deux types d'entités est donc difficile. Les machines intelligentes sont sensibles aux intrusions informatiques qui peuvent compromettre leurs "perceptions", la pertinence de leur "prise de décision" et leur capacité de gestion et de communication des données. Les humains ont d'autres faiblesses, comme la fatigue, la mémoire limitée et les capacités cognitives fragiles et facilement influencées. Dans ce contexte, une solution est de favoriser l'établissement de relations de surveillance de performance constructives entre les experts humains, entre les machines et, dans les deux sens, entre les experts et les machines [Desclaux 2021][Desclaux 2022].

Pour prendre en compte la multitude des situations liées à notre contexte applicatif et avancer sur cette notion de compréhension mutuelle entre opérateurs et systèmes artificiels, nous avons très tôt fait le choix d'orienter notre sujet d'étude vers une réflexion générale sur l'interaction entre l'Humain et les technologies issues de l'intelligence artificielle. Une étude qui pourrait être étiquetée de diverses manières, au croisement entre recherche fondamentale, philosophie des technologies, études des facteurs humains et contribution

pour l'informatique théorique. L'objectif est de formuler et de définir des concepts, hypothèses et principes qui serviront les acteurs de l'IA, qu'ils soient commanditaires industriels, concepteurs ou utilisateurs, à penser les systèmes d'aujourd'hui et de demain. Il s'agit donc bien d'une étude d'un TRL (technology readiness level) de bas niveau, comme la formulation de concepts technologiques (niveau 2 « Technology concept formulated ») voir la mise en forme de principes de base (niveau 1 « Basic principles observed »).

Pour utiliser les systèmes d'IA de manière appropriée, les experts devront disposer du savoir nécessaire à la compréhension du comportement des machines. Les utilisateurs devront posséder le savoir-faire requis pour utiliser les systèmes correctement. Les sociétés devront établir des normes et réglementations encadrant la conception et les usages des outils artificiellement intelligents.

Cependant, les lumières du savoir ne sauraient être uniquement portées sur la machine. De tout temps, l'Homme utilise l'outil et parfois s'y aliène. Imparfaites, les décisions et actions de l'individu peuvent entraîner, volontairement ou non, l'utilisation préjudiciable du plus inoffensif des objets. Pour correctement utiliser son outil, l'utilisateur doit intégrer les limites de sa propre intelligence, les biais qui affectent ses jugements et les fausses idées qui dévient ses raisonnements. L'utilisateur doit prendre en compte son propre fonctionnement pour garder un esprit critique, la souveraineté de ses décisions et la responsabilité de ses actes. Car l'outil ne reste qu'outil, quand bien même il serait intelligent.

Pour appréhender l'IA il faut appréhender l'intelligence naturelle. Pour comprendre comment utiliser correctement l'IA, il faut comprendre les comportements humains. Nous décelons ici une opportunité particulière. Celle de faire d'une pierre deux coups à travers la modélisation du fonctionnement de la cognition humaine. Ce modèle permettra d'une part de concevoir de meilleurs systèmes artificiellement intelligents, lorsque le modèle éclaire la conception et le fonctionnement de la machine. Et d'autre part, il permettra simultanément de mieux utiliser l'outil, lorsque le modèle éclaire les comportements humains des utilisateurs.

Une opinion communément admise argumente que de nombreux aspects de l'intelligence sont inintelligibles et irrémédiablement ambigus. Pour beaucoup, des notions telles que l'inconscient et les émotions, sont informelles et non modélisables. Pourtant, on sait maintenant que les processus non conscientisés de la psyché, l'expérience, la personnalité et les émotions sont des variables qui influencent grandement les comportements et les prises

de décisions de l'individu. On peut également affirmer que ces processus concourent aux performances de nos capacités cognitives et qu'ils sont indissociables de la notion d'intelligence.

A travers les civilisations et les époques, les idéologies religieuses, les écrits littéraires, les productions artistiques et les théories scientifiques ont retranscrit et utilisé les différents aspects de notre vie cognitive. A partir d'un formalisme des systèmes de traitement de l'information, nous proposons de retrouver ces différents savoirs et d'en discuter différemment les contenus.

Dans ce manuscrit, nous nous concentrons sur ce qui a trait aux sciences. Nous nous sommes intéressés et avons recoupé des théories et hypothèses provenant de différentes disciplines qui sont usuellement regroupées sous l'appellation de sciences cognitives.

Comme si nous menions une enquête, nous avons consulté les écrits d'auteurs classiques dans les domaines de la psychologie et de la philosophie de l'esprit. Nous avons exploré les théories cohérentes avec les observations modernes faites en neurosciences et en intelligence artificielle.

Cette approche, transdisciplinaire et générale, induit nécessairement la simplification d'éléments locaux. Un expert spécialisé pourra probablement discuter des généralisations que nous faisons sur les notions de son domaine. Pour autant, les raccourcis que nous prenons et la généralisation que nous en faisons sont nécessaires. Ils permettent de garder dans des proportions intelligibles les fondations théoriques que nous utilisons pour bâtir un édifice qui n'a de valeur que s'il est stable.

De ce fait, notre modélisation intègre nécessairement une part de subjectivité. Pour insister sur ce point, nous ne décrirons pas le fonctionnement de l'intelligence humaine. En lieu et place nous décrirons le fonctionnement de la cognition d'un système artificiel de traitement de l'information. Notre description et nos affirmations portent sur une machine formelle, hypothétique. La question de savoir si l'intelligence humaine fonctionne comme l'intelligence de la machine que nous décrivons est ouverte et chacun peut avoir son opinion à ce sujet.

Nous avons nommé l'architecture cognitive IPSEL, pour « Information Processing System with Emerging Logics ». Le nom que nous avons choisi fait référence à notre thèse principale. Dans cette thèse, nous soutenons que l'émergence de règles logiques modélisant

le dynamisme d'un ensemble de représentations est le principe qui permet à un système de traitement de l'information de mettre en œuvre ce qui s'apparente à une intelligence générale.

Dans ce manuscrit nous décrivons en détails la manière dont les composantes du modèle IPSEL interagissent et comment leur architecture permet, selon nous, la mise en œuvre d'une intelligence générale au sens humain : une intelligence parfois intuitive, d'autres fois raisonnable, une intelligence qui se constitue au cours de l'expérience et qui varie d'un individu à un autre, une intelligence émotionnelle et créative capable de concevoir des outils complexes comme peut l'être le langage.

Un modèle est toujours une approximation, une description simplifiée d'un sujet ou d'un phénomène. La spécificité de notre modèle tient à l'endroit où nous avons placé le curseur de la simplification. Notre volonté est de disposer d'un modèle suffisamment simplifié pour être utile aux commanditaires, concepteurs et utilisateurs de l'IA, qui ne possèdent pas forcément de connaissances approfondies en informatique ou en psychologie comportementale. Nous souhaitons dans le même temps disposer d'un modèle suffisamment détaillé pour permettre d'y intégrer et d'en discuter des aspects usuellement mis de côté dans la recherche en intelligence artificielle mais qui sont pourtant des constituants essentiels de l'intelligence naturelle. L'objectif que nous poursuivons est de remettre les émotions, l'expérience subjective et les comportements intuitifs au centre de la notion d'intelligence et de proposer un paradigme pour appréhender ce que pourrait être leur équivalent artificiel.

Le modèle théorique proposé dépeint avant tout une façon de penser l'intelligence. En guise de partie expérimentale, notre travail propose une forme d'implémentation des concepts élaborés en un pseudo-code qui montre la manière dont les différentes parties du modèle s'échangent l'information pour produire un comportement général. Notre temps a été utilisé pour affiner et détailler le modèle au maximum et proposer des exemples de son utilisation comme outil d'analyse. Pour confronter notre théorie à la réalité, nous proposons d'enrichir certains débats qui animent la communauté en IA, par exemple sur les notions d'IA de confiance, de créativité ou encore de conscience artificielle. La modélisation que nous proposons atteindra son but si elle permet effectivement aux personnes qui l'intègrent dans leur raisonnement, soit pour en suivre les principes, soit pour les réfuter, de mieux

appréhender la notion d'intelligence de niveau humain et de disposer de plus de clés de réflexion pour commanditer, spécifier, construire ou utiliser l'intelligence artificielle.

Chapitre 2 : Etat de l'Art

Dans ce chapitre nous présentons les différentes études bibliographiques que nous avons menées. Portant sur plusieurs disciplines, ces études nous ont permis de cadrer le périmètre de notre recherche. Il s'est s'agit de faire un état des lieux de ce à quoi correspond la confiance humaine dans l'utilisation d'un outil automatisé, d'explorer le domaine de l'intelligence artificielle pour en comprendre les technologies qui en sont issues, de prendre connaissance des différentes idées et théories que la communauté scientifique a formulées à propos de la cognition humaine et enfin de s'inspirer du formalisme et des méthodes employées en architecture cognitive, une sous-discipline de l'IA qui cherche à modéliser des systèmes cognitifs artificiels.

2.1 Confiance Homme/Automation

La confiance de l'Homme envers un outil automatisé constitue un véritable domaine de recherche. Nous proposons ici une synthèse destinée à discuter les hypothèses, connaissances et théories qui ont été proposées dans un ensemble de ressources que nous avons sélectionnées. L'idée principale que nous cherchons à dégager pourrait synthétiquement se formuler comme suit : les études concernant le sujet de la confiance envers un outil automatisé ont commencé dans les années 70 à la suite de l'arrivée de l'automatisation dans les processus industriels ; elles tendent à montrer que la confiance est un état physio-psychologique de l'individu qui détermine sa propension à utiliser l'outil dans un contexte caractérisé par l'incertitude et le risque.

2.1.1 La confiance comme sujet de recherche

Le thème de la confiance est un sujet de recherche étudié par différentes disciplines comme la psychologie, la sociologie et les sciences économiques. Plusieurs définitions ont émané de ces recherches qui se sont développées à partir des années 80.

Les définitions portées sur la confiance varient en fonction des auteurs, de la discipline, du contexte et de la langue utilisée. Il ne semble pas exister de définition générale et universelle qui serait consensuelle, acceptée et utilisée par toutes les disciplines scientifiques qui s'y intéressent [Rousseau 1998].

Nous reprenons ici quelques-unes de ces définitions provenant de publications scientifiques que nous avons sélectionnées au vu du nombre de références dont elles ont fait l'objet (référencements relevés sur google scholar).

« La confiance dans le fait que l'on trouvera ce qu'on attend d'un autre, plutôt que ce qu'on craint ». [Deutsch 1973], *7800 référencements*.

« Attente liée à la probabilité subjective qu'un individu attribue à la survenance d'un ensemble d'événements futurs ». [Rempel et al. 1985] *5100 référencements*.

« La confiance est l'attente, détenue par un membre d'un système, de la persistance des ordres sociaux, naturels et moraux, d'une performance technique et de la responsabilité fiduciaire d'un membre du système, et est liée à, mais pas nécessairement isomorphe avec des mesures objectives de ces qualités. ». [Muir 1987] *980 référencements*.

« Volonté d'une partie d'être vulnérable aux actions d'une autre partie sur la base de l'attente que l'autre effectuera une action particulière importante pour le donneur de confiance, indépendamment de la capacité de surveiller ou de contrôler cette partie ». [Mayer 1995] *28000 référencements*.

« Un état psychologique comprenant l'intention d'accepter la vulnérabilité basée sur des attentes positives des intentions ou du comportement d'un autre ». [Rousseau 1998] *13800 référencements*.

« Un état de vulnérabilité ou de risque perçu et qui découle de l'incertitude d'un individu concernant les motivations, les intentions et les actions d'un autre individu dont il dépend ». [Kramer 1999] *5000 référencements*.

« L'attitude selon laquelle un agent aidera à atteindre les objectifs d'un individu dans une situation caractérisée par l'incertitude et la vulnérabilité ». [Lee et al. 2004] *3800 référencements*.

On peut constater que les définitions diffèrent en fonction des sujets étudiés. Il existe plusieurs types de confiance comme par exemple la confiance en soi, la confiance interpersonnelle, la confiance dans les institutions ou encore la confiance envers un outil. Plus particulièrement, les études pertinentes par rapport à notre sujet vont dans le sens de la définition de la confiance envers un outil automatisé. L'automatisation est définie comme la

réalisation automatique d'une tâche précédemment réalisée par l'homme [Parasuraman 1997]. Ainsi on peut émettre l'hypothèse que l'intelligence artificielle est une forme d'automatisation portant sur des processus que l'on attribue généralement à la cognition humaine.

Dans notre travail, la définition que nous retenons pour la confiance est la suivante : L'état physico-psychologique de confiance est généré par la cognition de l'homme qui procède à une évaluation consciente et inconsciente de la probabilité que le système accomplisse correctement la tâche qui lui est attribuée et l'aide ainsi à atteindre son objectif.

2.1.2 Le rôle de la confiance dans l'interaction Homme-Automation

Les différentes analyses et expérimentations menées dans le cadre de la confiance Homme-Automation tendent à montrer le rôle de la confiance en tant que variable définissant la propension d'un opérateur à correctement utiliser un système automatisé [Barber 1983][Endsley 1987][Zuboff 1988].

Cette théorie trouve écho dans les sciences cognitives qui définissent la confiance comme un état émotionnel servant à orienter l'individu vers un comportement d'approche ou de fuite [Kahneman 2011]. Un défaut de confiance orienterait l'individu vers une sous-utilisation de la technologie, alors qu'un excès de confiance mènerait à une utilisation irraisonnée (« surutilisation ») de ladite technologie.

Le processus d'évaluation de la confiance attribuable à une technologie automatisée, que l'on appelle « calibration de la confiance », est une étape critique pour la bonne utilisation de cette dernière. Permettre à l'homme de calibrer correctement la confiance qu'il peut attribuer aux technologies de l'intelligence artificielle est donc un champ de recherche de grande importance.

De ces études réalisées sur la confiance envers l'automation ressortent certains éléments communs que l'on retrouve dans la majorité des définitions proposées pour la confiance. Nous résumons ces caractéristiques qui semblent faire consensus en listant les points suivants :

- ➔ La confiance est un état psychologique.
- ➔ La confiance représente une espérance quant à un événement futur.
- ➔ La confiance est une évaluation subjective.

➔ La confiance implique l'acceptation d'un niveau de risque et de vulnérabilité.
➔ La confiance oscille entre deux états asymptotiques que sont la défiance et l'excès de confiance en passant par des états intermédiaires comme la méfiance.
➔ La confiance est facile à perdre, difficile à gagner, et encore plus difficile à retrouver une fois qu'elle a été perdue.

➔ La confiance se propage vers les divers sous-ensembles d'un système.

Nous proposons alors d'établir la définition suivante, qui reprend de manière synthétique les points relevés ci-dessus.

Confiance envers un système artificiel : État physio-psychologique de l'utilisateur, engendré par l'évaluation de la possibilité que le système puisse l'aider à accomplir ses objectifs, dans une situation caractérisée par l'incertitude et la vulnérabilité.

L'état physico-psychologique de confiance est généré par la cognition de l'homme qui procède à une évaluation consciente et inconsciente de la probabilité que le système accomplisse correctement la tâche qui lui est attribuée. Pour apporter des hypothèses utiles quant à la manière de permettre aux utilisateurs de correctement calibrer la confiance qu'ils attribuent aux systèmes d'intelligence artificielle, il nous faudra donc essayer d'avancer vers une certaine compréhension des mécanismes cognitifs de l'homme, ainsi que vers une certaine compréhension de ce que sont les systèmes d'intelligence artificielle.

2.2 Domaine de l'Intelligence Artificielle

L'intelligence artificielle est un domaine de recherche né dans les années 50 sous l'impulsion d'un groupe de chercheurs de divers horizons. Son objectif est de formaliser les mécanismes cognitifs de l'Homme en termes de processus de traitement de l'information, pour qu'ils puissent ensuite être réalisés par des machines. Son horizon est la conception d'une machine artificielle qui déploierait les mêmes capacités de raisonnement, de communication et d'intelligence que l'Homme. Cet horizon est pour ainsi dire le « graal » des chercheurs en IA.

Nous présentons le domaine de l'intelligence artificielle d'un point de vue général, en relatant de manière succincte son histoire, ses accomplissements, ses problématiques actuelles et les directions prises par la communauté pour les résoudre. L'objectif n'est pas de

fournir un état de l'art détaillé des différents sous-domaines qui constituent l'intelligence artificielle. Il s'agit plutôt de prendre de la hauteur sur ses concepts et ses productions.

Nous procédons d'abord à une rapide présentation de l'histoire du domaine. Puis nous discutons des tentatives de formalisation de la notion d'« intelligence » avant de discuter les techniques algorithmiques qui cherchent à la produire. Ces techniques sont usuellement regroupées en deux courants ; l'approche connexionniste et l'approche symbolique. Les systèmes qui les implémentent peuvent, quant à eux, être distingués en deux catégories. Les systèmes dont la fonction est de simuler une partie de l'intelligence humaine pour être appliqués à la réalisation d'une tâche spécifique dans un domaine restreint, que l'on nomme alors « Artificial Narrow Intelligence ». Et les systèmes qui ont vocation à mettre en œuvre l'ensemble des fonctions cognitives humaines pour évoluer et s'adapter à des environnements complexes et variés. Ce second type de système est dénommé « Artificial General Intelligence ». Nous présentons ensuite rapidement les principales réalisations notables du domaine ainsi que celles relatives aux secteurs de la défense. Pour finir ce tour d'horizon du domaine de l'intelligence artificielle nous présentons l'une de ses disciplines : l'Architecture cognitive, qui s'attache à modéliser la manière dont les techniques de l'IA devraient se combiner les unes aux autres pour atteindre le graal, l'intelligence artificielle général, celle de niveau humain. Enfin, nous présentons l'état d'avancée des travaux menés dans le cadre de l'IA de confiance.

2.2.1 Histoire du domaine de l'Intelligence Artificielle

Le behaviorisme considère la cognition comme une boîte noire. Dans ce paradigme il est impossible, au sein de la boîte, de modéliser les composants qui restent inconnus et inaccessibles. A contrepied de ce courant de pensée, se met en place au début du 20 -ème siècle la démarche constructiviste. Rudolf Carnap voit en elle le moyen d'avancer vers un monde « inter subjectivement objectif ». Il écrit en 1928 dans son essai « The logical structure of the world » : « Même si l'origine subjective de toutes les connaissances réside dans le contenu des expériences et de leurs relations, il est toujours possible, comme le montrera le système constructiviste, de progresser vers un monde objectif, intersubjectif, qui peut être compris conceptuellement et qui est identique pour tous les observateurs » [Carnap 1928]. Carnap et d'autres de ses collaborateurs du cercle de Vienne comme Bertrand Russel ou Ludwig Wittgenstein ouvrent alors la boîte de pandore et influenceront

la façon de penser d'une nouvelle génération de scientifiques à travers des textes fondateurs comme le « Principia Mathematica » [Whitehead & Russell 1910] .

Ces figures de la nouvelle génération se nomment Von Neumann, Turing, Wiener etc. Au milieu des années 50, une fois sortis des bancs des prestigieuses universités anglo-saxonnes, ils s'essaient chacun à leur manière à la formalisation systémique des processus de traitement de l'information observés chez l'être Humain, dans l'espoir de les recréer artificiellement. Réalisées pour des finalités très pragmatiques, comme la conception de systèmes de guidage des missiles anti-aériens ou bien le décryptage des communications allemandes durant la seconde guerre mondiale, ces réflexions ont fait naître l'ordinateur comme on le connaît aujourd'hui. On doit à John Von Neumann l'architecture de ses composants [Neumann 1945] [Neumann 1948], à Norbert Wiener les processus de rétrocontrôle et le domaine qui les étudie – la cybernétique - [Wiener 1948], et à Alan Turing la description du mécanisme élémentaire de ces machines : La machine de Turing [Turing 1950].

Ils ne sont pas les seuls à s'intéresser à ces sujets et en 1956, lors d'une série de conférences à l'université Dartmouth College, plusieurs scientifiques venant de différentes disciplines, académiques et industrielles, se rencontrent pour discuter de la thèse selon laquelle « chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peut être si précisément décrit qu'une machine peut être conçue pour le simuler » [Crevier 1993]. Les participants à cette rencontre historique, Marvin Minsky, John Mc Carthy, Claude Shannon, Nathan Rochester, Ray Solomonoff, Oliver Selfridge, Trenchard More, Arthur Samuel, Allan Newell et Herbert Simon s'accordent autour de la proposition de Mc Carthy d'intituler ce nouveau champ de recherche « Intelligence Artificielle ». Cet événement marque le moment fondateur de l'intelligence artificielle en tant que discipline théorique indépendante de l'informatique.

Après des dizaines d'années de développement, les conférences en intelligence artificielle ne rassemblent plus quelques dizaines de scientifiques comme à Dartmouth, mais des dizaines de milliers. Le domaine est même subdivisé en sous-domaines, chacun référant à un aspect particulier du traitement intelligent de l'information et à ses applications. Vision par ordinateur, traitement automatique du langage naturel, apprentissage statistique, résolution de problème par contraintes, arbres de recherche, neurones artificiels etc. sont autant de disciplines spécifiques qui ont éclos des recherches en IA. Bien entendu l'entreprise ne fut pas sans encombre. L'intérêt et les investissements attribués aux domaines

ont oscillé au cours de l'histoire en fonction des contextes, des courants de pensée, des résultats. Tantôt porté par l'engouement des perspectives et des nouvelles découvertes, tantôt profondément remis en question par les déceptions relatives aux performances, ou plutôt aux non-performances, de systèmes trop souvent idéalisés par leurs concepteurs. L'histoire de l'IA est symboliquement perçue comme la succession « d'étés et d'hivers », illustrant ces oscillations entre promesses théoriques et réalités applicatives. Nous ne parcourons pas cette histoire en détail, convaincu que de nombreux livres ont été écrits sur ce sujet et suggérons le récit de Daniel Crevier « AI : The tumultuous search for Artificial Intelligence » [Crevier 1993].

Malgré l'étendue actuelle du domaine et des nombreuses disciplines qui le composent, l'IA est un sujet de recherche encore jeune. Les programmes développés dans son cadre sont nombreux et en constante évolution. Les fonctions que doivent accomplir ses programmes sont diverses. La seule certitude concerne son objectif : concevoir des machines dotées d'une intelligence comparable à celle des humains, une intelligence artificielle.

2.2.2 Formalisation du concept d'intelligence

La promesse du domaine de l'intelligence artificielle, exprimée explicitement dans les années 50 et amplement répétée depuis, est de développer des systèmes dotés d'une intelligence comparable à celle de l'homme. Il semble donc nécessaire, dans un premier temps, de définir l'intelligence.

Il existe des définitions de sens commun que l'on trouve dans les dictionnaires, mais elles ne sont pas utiles d'un point de vue informatique dans le sens où elles sont ambiguës et ne donnent pas de critère de mise en œuvre et de mesurabilité dans le cas d'une machine. En réaction, il a souvent été de mise de décrire non pas des définitions, mais des tests cherchant à la mesurer, comme le Test de Turing et nombre de ses dérivés. Pour François Chollet, ce type de test n'est pas pertinent car il se soustrait à une définition formelle et objective de l'intelligence, et relègue la tâche de mesure à un juge Humain qui lui-même à une conceptualisation subjective de la notion d'intelligence [Chollet 2019]. Pour lui, l'incapacité des acteurs du domaine de l'IA à répondre de manière satisfaisante à la question de ce que la discipline entend par « intelligence » est une preuve de son immaturité. Qui plus est, très peu de recherches sont consacrées à progresser sur ce point. Ce qui peut être expliqué par le fait qu'elles ne présentent que très peu d'intérêts commerciaux.

En 2007, les chercheurs Shane Legg et Marcus Hutter concluent l'un des rares rapports sur les définitions et méthodes de mesures formelles de l'intelligence, par la phrase « *to the best of our knowledge, no general survey of tests and definitions has been published* » [Legg & Hutter 2007]. Une constatation réitérée en 2017 par Hernandez Orallo [Hernandez 2017]. Hutter et Legg ont bien essayé de fournir une définition formelle dans leur publication « Universal intelligence : A definition of machine intelligence » [Legg & Hutter 2007-2], mais elle n'a pour le moment pas trouvé de réelle application car elle est jugée trop générale et incalculable. Pour ces auteurs, l'intelligence mesure la capacité d'un agent à accomplir des objectifs dans un vaste domaine d'environnement et réfère à la théorie de l'« Universal optimal learning agents ». Cette définition est difficile à mettre en œuvre mais on y retrouve une notion qui est souvent associée à l'intelligence : l'apprentissage.

Étymologiquement, le mot intelligence provient du latin « intelligentia » qui signifie « faculté de comprendre ». Il est dérivé du mot « intelligere » qui signifie « choisir entre ». On retrouve là les aspects qui sont couramment utilisés dans les définitions provenant des sciences humaines : l'acte d'intégrer des connaissances et de s'en servir pour décider entre plusieurs actions. Par exemple Alfred Binet définit l'intelligence comme la faculté d'un individu à s'adapter aux circonstances [Binet & Simon 1916].

Cette définition est loin d'être absolue et encore moins formelle. Elle nous permet juste de mettre en lumière trois aspects fondamentaux des techniques et systèmes conçus dans le cadre de l'intelligence artificielle :

1. La gestion des connaissances
2. Les méthodes d'apprentissages
3. Les domaines d'adaptation

En effet, les méthodes d'obtention et de représentation des connaissances, les méthodes d'apprentissage et les domaines auxquels les systèmes doivent s'adapter sont trois problématiques qui servent souvent de points d'ancrage pour distinguer les approches et techniques de l'intelligence artificielle. Relativement aux modes d'obtention et de représentation des connaissances, deux approches s'opposent historiquement : le connexionnisme et le cognitivisme (ou approche symbolique). Les méthodes d'apprentissage sont couramment subdivisées en apprentissage supervisé, non supervisé et

associatif. Enfin l'adaptation des systèmes peut être considérée à deux niveaux : Les « Artificial Narrow Intelligence » qui doivent s'adapter à un domaine restreint, et les « Artificial General Intelligence » qui ont pour raison d'être l'adaptation dans toute sa généralité.

2.2.3 Approche connexionniste et approche symbolique

De nombreux auteurs se sont employés à modéliser les différents mécanismes de la cognition humaine. Dans cette optique, le neurone artificiel a été introduit par les chercheurs Warren S. McCulloch et Walter Pitts [McCulloch & Pitts 1943] et a permis la conception du perceptron [Rosenblatt 1958], puis d'autres modèles d'architecture de réseaux de neurones [Hopfield 1982] et des réseaux multicouches [LeCun & Bengio 1995], grâce au mécanisme d'apprentissage tiré du principe de backpropagation [Rumelhart 1988]. Parallèlement à cette approche computationnelle où le modèle est dirigé par les données « data-driven models » qui se verra dénommé « connexionnisme », une autre partie de la communauté informatique a théorisé et expérimenté des systèmes à base de règles logiques « rules-based models » en suivant une approche dite « symbolique » car prenant la forme de manipulation de symboles. L'aboutissement de ces travaux inaugurés par le « general problem solver » d'Alan Newell et du prix Nobel Herbert Simon [Newell & Simon 1959] mèneront aux technologies des « systèmes experts ». Les deux approches ont tour à tour eu leurs moments de gloire, et bien que les réseaux de neurones soient prépondérants dans les recherches actuelles, les hybridations de ces deux voies sont essentielles dans la conception de systèmes complexes.

Depuis lors deux écoles de pensée se disputent pour savoir laquelle de ces approches, le connexionnisme (surnommé « scruffies ») ou le symbolisme (surnommé « neats »), est la plus féconde. Le connexionnisme se veut biomimétique, il s'inspire des neurones biologiques, de l'auto-apprentissage et est très performant pour la simulation des actions sensori-motrices. Les techniques incarnant cette approche sont celles du Machine Learning, du Deep Learning et des réseaux de neurones. Le symbolisme quant à lui ne cherche pas le biomimétisme, il modélise le raisonnement logique et représente les connaissances sous forme d'objets et de symboles qu'il associe entre eux par des règles logiques. On y trouve des techniques d'inférences logiques, de manipulation de symboles, d'arbre de recherche ou de programmation par contraintes.

Pour Pedro Domingos [Domingos 2015] il existerait cinq grands courants dans l'IA : le symbolisme, le connexionnisme, les méthodes évolutionnistes, les méthodes probabilistes telles que les réseaux bayésiens et les analogistes utilisant des méthodes de « clustering ».

De notre point de vue, ces distinctions entre les approches ne sont pas pertinentes et s'apparentent à des oppositions « de chapelles ». Les systèmes efficaces ont généralement toujours recours à plusieurs de ces techniques et il est rare de pouvoir définir un programme comme purement connexionniste ou symbolique. Notons aussi que ces considérations résultent peut-être d'un positionnement subjectif par rapport à l'ordre de grandeur du référentiel employé. Tout programme informatique utilise des symboles et des règles, les nombres encodant les poids des réseaux de neurones sont des symboles, les fonctions d'activation et de rétro-propagation utilisent des règles mathématiques, les statistiques sont des règles appliquées à des symboles. Un neurone artificiel pris de manière isolée se comporte comme un système expert qui ne serait doté que d'une seule règle. Enfin le caractère collectif du connexionnisme se retrouve dans le symbolisme qui définit des objets comme combinaisons de symboles desquelles émerge un sens. Un autre argument réside dans le fait que les méthodes ainsi nommées ne cherchent pas à accomplir les mêmes fonctions, leur comparaison est donc illusoire. Le connexionnisme cherche à simuler les fonctions sensori-motrices de l'homme : la reconnaissance de forme, le contrôle fluide et automatisé. Le symbolisme en revanche cherche à simuler le raisonnement conceptuel. Ces deux types de fonctions sont distinctes chez l'homme (cf. partie sur la dual-process theory) et il pourrait sembler normal que leurs accomplissements passent par des méthodes distinctes et incomparables.

Peut-être est-il plus utile de différencier ces approches en fonction du type d'apprentissage qu'elles utilisent. L'apprentissage connexionniste utilise des données provenant de l'environnement selon un mode de raisonnement inductif alors que le symbolisme utilise des règles qui lui ont été données et les utilise via un mode de raisonnement déductif. Cependant, une fois de plus, la distinction est discutable puisque nombre de systèmes symboliques induisent des règles à partir de données de l'environnement, et les programmes connexionnistes ne sauraient fonctionner sans un minimum de règles mathématiques et statistiques qui leurs sont intégrées préalablement.

En conclusion, cette dichotomie entre connexionnisme et symbolisme provient du positionnement historique entre deux communautés de recherches ayant abordé une même

problématique selon différents angles. De nos jours, les systèmes d'IA font intervenir simultanément des techniques provenant des deux tendances. Les récentes directions prises par la communauté, nommée hybridation ou calcul neuro-symbolique, incarnent l'obsolescence d'un tel débat. Ces méthodologies consistant à combiner les deux approches ont gagné une attention particulière ces dernières années et certains auteurs les qualifient de « troisième vague de l'IA » [Garcez & Lamb 2020].

2.2.4 La dual process theory en intelligence artificielle

La communauté s'accorde à dire que certaines caractéristiques de l'intelligence humaine font toujours défaut aux systèmes d'IA actuels. Elles sont nommées adaptabilité, généralisation, sens commun, raisonnement de causalité, robustesse, explicabilité ou encore raisonnement éthique. Pour Yoshua Bengio il est nécessaire que les systèmes généralisent à partir de données brutes pour induire un flux de concepts reliés entre eux [Bengio 2017]. Gary Marcus argumente que la manipulation de connaissances explicites et de symboles est un prérequis pour le développement de systèmes robustes [Marcus 2020]. L'un des principaux débats qui anime à ce jour le domaine est de savoir si un réseau de neurones artificiels de bout en bout (« end to end neural network ») peut atteindre ces capacités manquantes ou bien s'il est nécessaire de combiner (par hybridation) les approches de Machines Learning avec des techniques basées sur la logique et les symboles [Booch et al. 2020]. Les recherches menées dans le sens de l'hybridation sont regroupées derrière le terme de calcul neuro-symbolique (« Neuro-symbolique computation ») [Garcez & Besold 2019]. Pour avancer sur ces questions, la communauté de l'intelligence artificielle s'inspire et exploite les théories cognitives, en particulier la théorie du double processus (dual-process theory), démocratisée avec les travaux de Kahneman (voir à ce sujet la partie sur la cognition humaine), qui infuse les recherches en IA depuis très récemment [Booch 2020]. Un parallèle est généralement fait entre système 1/système 2 et les deux approches courantes de l'IA. Le système 1 semble avoir des caractéristiques similaires à celles des techniques de Machine Learning alors que les processus du système 2 ressemblent à ceux étudiés dans le cadre de l'IA symbolique et du raisonnement logique. Cependant la comparaison ne peut être totale. En science cognitive le système 1 a la capacité de construire des relations de causalité basique et de raisonner par sens commun, des compétences qui font défaut aux techniques de Machines Learning [Marcus 2020]. De plus, l'activité du système 2 est dite séquentielle et requiert une attention totale de la part du sujet, ce qui ne correspond à rien dans le cas d'une machine [Booch et al. 2020].

Cette inspiration provenant des sciences cognitives est productive et alimente les recherches portant sur l'hybridation des deux approches de l'intelligence artificielle et sur le calcul neuro-symbolique. Elle permet de mettre en avant des questionnements relatifs à la manière de penser une forme de dual-process theory pour les machines. [Booch et al. 2020] propose de les formuler en 10 points qui devraient selon lui constituer des axes de recherches à investir :

1. Est-il nécessaire de clairement identifier les capacités du système 1 et du système 2 dans le cas de machines artificielles ? Quelles en seraient alors les caractéristiques ?

2. L'aspect séquentiel du système 2 est-il un « bug » ou une caractéristique ? Doit-il être intégré aux machines ou est-il possible d'utiliser le calcul parallèle pour les processus de raisonnement d'un système 2 artificiel ?

3. Quelles sont les métriques permettant d'évaluer la qualité d'un système hybridant les deux approches système 1 / système 2 (Machine Learning / Symbolic reasoning) ?

4. Comment définir l'introspection en termes d'IA et comment l'introspection est-elle liée à l'autonomie et à la coopération Homme-Machine ?

5. Comment modéliser la gouvernance des systèmes 1 et 2 ? Quand et comment devons-nous décider quel système doit agir et comment leur coopération se fait-elle ?

6. Comment pouvons-nous tirer parti d'un modèle basé sur le système 1 et le système 2 dans l'IA pour comprendre et raisonner dans des environnements complexes lorsque nous avons des priorités concurrentes ? Quelle partie du système d'IA reconnaît la divergence et, soutenu par l'introspection, identifie la meilleure procédure de résolution ?

7. Quelles capacités sont nécessaires pour effectuer diverses formes de jugement moral et de prise de décision ? Comment modélisons-nous et déployons-nous des théories éthiques normatives potentiellement contradictoires dans l'IA ?

8. Comment définir des mécanismes d'abstraction / généralisation guidés par une notion d'attention et passant du niveau des données brutes à un niveau plus abstrait ? Comment savons-nous ce qu'il faut oublier des données d'entrée pendant l'étape d'abstraction? Devrions-nous garder les connaissances à différents niveaux d'abstraction, ou simplement des données brutes et des connaissances de haut niveau pleinement explicites ? Qu'est-ce

que cela signifie que la connaissance soit explicite : est-ce lié à la présence de métadonnées, de graphiques de connaissances structurés ou bien d'entités liées au langage ?

9. Dans une vue multi-agent de plusieurs systèmes communiquant et apprenant les uns des autres, comment exploiter / adapter les résultats actuels sur le raisonnement épistémique et comment construire / apprendre des modèles de l'environnement et des autres agents ?

10. Quels sont les meilleurs choix architecturaux pour soutenir la vision ci-dessus de l'avenir de l'IA ?

Ces questions se rapportent à la conception d'un système mettant en œuvre les principes de la théorie psychologique du Dual-Process. Il semble donc que la finalité d'un tel système soit de mettre en marche l'ensemble des capacités cognitives de l'Homme. En Intelligence Artificielle, un sous-domaine de recherche s'intéresse tout particulièrement à ce type de finalité, il s'agit de l'architecture cognitive.

2.3 Cognition humaine

Cette partie présente les recherches bibliographiques que nous avons menées sur le fonctionnement de la cognition humaine. Ce que nous retenons de ces lectures pourrait synthétiquement se formuler comme suit: La cognition humaine est une notion complexe qui n'est pas parfaitement comprise. Son étude se scinde en plusieurs disciplines comme la linguistique, la psychologie, la philosophie, la neurologie, la sociologie que l'on regroupe parfois derrière le terme de sciences cognitives. Cependant, une idée semble émerger et a été formulée sous différentes formes au cours des siècles et dans différentes cultures. Cette idée consiste à dissocier deux entités collaborant pour réaliser la cognition. Parfois appelées l'inconscient et la conscience, le corps et l'esprit, l'intuition et l'entendement, les émotions et la raison, la théorisation de cette idée a pris le nom de « Dual-Process Theory ».

Pour écrire cette partie nous avons mobilisé des articles scientifiques concernant les sciences cognitives et la « dual-process theory ». Nous avons également consulté de nombreux livres de psychologie et de philosophie, notamment d'auteurs illustres comme Platon, Kant, Spinoza, Nietzsche, Freud etc.

2.3.1 Cognition

Etymologiquement le terme « cognition » est l'association de deux mots latins, con (« avec ») et gnosco (« connaître »). Le dictionnaire Larousse définit la cognition comme l'ensemble des structures et activités psychologiques dont la fonction est la connaissance. Ces activités mettent en jeu la mémoire, le langage, le raisonnement, l'apprentissage, l'intelligence, la résolution de problèmes, la prise de décision, la perception ou l'attention. Traditionnellement mises en opposition avec la notion d'affect, il est désormais communément accepté que les émotions, l'expérience et la personnalité jouent un rôle important dans l'accomplissement des fonctions cognitives [Damasio 2006] et qu'elles représentent une certaine forme de connaissance du sujet. Dans ce document, nous utilisons le terme cognition de manière étendue. La cognition désigne alors les processus de traitement de l'information dits de « haut niveau » mais aussi des processus plus élémentaires comme la perception, la motricité et les émotions [Wikipedia 1]. L'appareil cognitif est considéré comme un système complexe qui met le corps en mouvement en effectuant divers traitements faisant le lien entre perceptions, affects, pensées et comportements. Il s'agit de l'ensemble des facultés mentales associées à l'activité du système nerveux incluant celles qui sont plus automatiques. La cognition peut également être étendue à tout système, humain, animal ou artificiel capable d'acquérir, faire émerger du sens, conserver, utiliser ou transmettre des connaissances de manière à satisfaire une contrainte de viabilité. Pour certains auteurs il est possible de considérer que « cognition = vie » et que toute activité au sein d'un organisme vivant relève d'un phénomène cognitif avec une prise de connaissance et des boucles de rétroactions afin de maintenir l'ensemble des homéostasies (stabilisation et réglages des caractéristiques physiologiques) et participer aux processus autopoïétiques (propriété d'un système de se produire lui-même en interaction avec son environnement) [Stewart et al. 2010].

Les sciences cognitives peuvent être considérées comme un ensemble de six disciplines scientifiques que sont les neurosciences, la linguistique, l'anthropologie, la psychologie, la philosophie de l'esprit et l'intelligence artificielle. Ces domaines de recherche s'intéressent tous à la cognition, naturelle ou artificielle, ou bien à ses productions. Les sciences cognitives sont pluridisciplinaires par construction et mettent en place des interdisciplinarités permettant de faire émerger de nouvelles problématiques à la frontière de ces disciplines. Elles sont transdisciplinaires quand elles instaurent des paradigmes unificateurs, comme par exemple le concept de traitement de l'information. Prenant le

contre-pied du béhaviorisme, elles proposent de fournir des théories de ce qui se passe à l'intérieur d'un organisme, doté de capacités cognitives, quand il réalise un comportement. La « boîte noire » prenant place entre les stimuli et le comportement, proposée par le béhaviorisme, est remplacée par le concept de représentation. Cette notion de représentation, son étendue et ce qu'elle recouvre, est cependant encore très débattue [Steiner 2011]. Il existe historiquement trois paradigmes principaux au sein des sciences cognitives : l'approche symbolique computationnelle, l'approche connexionniste et dynamique, et l'approche incarnée et énactive. Se rajoutent à celles-ci des approches hétérodoxes comme l'approche bayésienne et les approches regroupées sous le sigle « 4EA » pour « Embodied, Embedded, Enacted, Extended, Affective » (Incarné, Situé, Enacté, Étendu, Affectif) [Monier 2019].

Dans l'approche symbolique computationnelle, l'esprit traite littéralement de l'information dans le sens où il manipule des symboles discrets selon un ensemble de règles [Fodor 1986]. Pour le connexionnisme la cognition n'est pas un système de règles de manipulation de symboles, mais un système d'entités élémentaires en interaction les unes avec les autres pour former des réseaux qui donnent naissance à des comportements dynamiques. Le connexionnisme introduit la notion d'émergence, ou d'auto-organisation, structurant des entités de haut niveau, à partir de l'interaction des entités de bas niveau. Dans l'approche incarnée et énactive, les phénomènes d'émergence et d'auto-organisation dynamiques sont aussi très importants, mais il y a une remise en cause de la pertinence des représentations comme éléments explicatifs pour la cognition. Le terme « énaction » a été inventé par Varela, Thompson et Rosch dans leur ouvrage « L'inscription corporelle de l'esprit » [Varela et al. 1993]. Les approches énactivistes considèrent l'esprit et la cognition comme résultant de l'interaction dynamique entre l'organisme vivant et son environnement, tel qu'il se déploie à travers les activités et les expériences de corps de vie situées en interaction réciproque. L'énactivisme, ainsi que les paradigmes théoriques connexes regroupés sous le sigle « 4EA », souligne qu'il y a une profonde continuité entre le mental et la vie définie comme autopoïétiques. Le « A » de « 4EA » renvoie aux affects, aux émotions et aux sentiments qui signalent l'état vécu de l'organisme, soit positif, associé à la joie, soit négatif, associé à la tristesse, et permettent ainsi d'avoir un système de valeurs directement associé à l'état homéostatique de l'organisme. Ainsi pour ces approches les affects jouent un rôle essentiel dans la relation au corps et sont totalement indissociables de la cognition [Damasio 2017].

Ces différentes approches ne sont pas nécessairement contradictoires et mutuellement exclusives mais constituent plusieurs niveaux de compréhension de la cognition, de son fonctionnement et de ses origines. Elles ont été développées à partir des études et théories provenant de la psychologie et, à notre connaissance, la plupart des idées qu'elles portent ont été initialement proposées en philosophie, avant même que la psychologie n'existe en tant que science. Par exemple, la vision des systèmes vivants comme structure biologique ne s'attachant fondamentalement qu'à leur propre conservation, correspond à la vision protobiologique du conatus de Spinoza [Damasio 2003][Monier 2019][Jaquet 2009].

La psychologie est issue de deux autres disciplines, la physiologie et la philosophie. On attribue sa fondation à Wilhelm Wundt [Wundt 1904]. Un autre de ses fondateurs, William James, la définit comme la science de la vie de l'esprit. Si la philosophie pouvait s'apparenter à de la spéculation, la psychologie se veut basée sur l'étude scientifique de la biologie du cerveau. Pour Steven Pinker, la personnalité et l'intelligence sont probablement déjà programmées depuis la vie utérine et certainement influencées par des facteurs hormonaux [Pinker 2003]. Les comportements, ou du moins leurs tendances, sont également déterminés par des facteurs biologiques [Brizendine 2006][Moir & Jessel 1991]. De cette biologie du cerveau émerge l'esprit et les phénomènes psychologiques. Pour Olivier Sacks, le cerveau travaille constamment à créer et à entretenir la perception d'un « je » et un « sentiment du moi » [Sacks 2018]. Malgré l'empreinte dominante de la biologie sur nos comportements, les facteurs environnementaux sont déterminants quant à leurs expressions particulières, comme l'exposent les travaux de Jean Piaget [Piaget et al. 1925] et Alfred Kinsey [Kinsey et al. 2003]. Les phénomènes psychologiques et les comportements de l'homme doivent donc être considérés comme résultant d'une combinaison entre nature et culture, ils sont conditionnés par l'ADN, les hormones, l'architecture du cerveau et par des facteurs environnementaux. La psychologie concerne également l'inconscient. Un terme popularisé dans le domaine par les travaux de Sigmund Freud [Freud et al. 1954] et de Carl Jung [Jung 1964]. Encore une fois ce concept d'inconscient avait déjà été présenté, sous d'autres appellations, en philosophie notamment par Arthur Schopenhauer [Schopenhauer 1888] et Friedrich Nietzsche [Nietzsche 1968].

C'est dans ce contexte des études sur la cognition que nous proposons de pointer plus en détail la théorie qui nous semble la plus prometteuse pour débattre de la mécanisation de la prise de décision : la théorie des processus duaux (ou dual-process theory), théorie qui

occupe actuellement une place de choix dans les débats en psychologie cognitive et en philosophie de l'esprit.

2.3.2 La *Dual Process Theory*

Durant les trois dernières décennies, de nombreuses recherches en psychologie cognitive ont été guidées par une classe de théories généralement décrites sous l'appellation de « Dual Process Theories » [Gawronski & Creighton 2013]. La caractéristique commune de ces théories est qu'elles divisent les processus mentaux responsables des jugements et des comportements en deux catégories générales ; les processus qui opèrent automatiquement et les processus qui opèrent de manière contrôlée.

Les premiers travaux sur ce sujet se sont concentrés sur des phénomènes spécifiques comme la persuasion (exemple : [Chaiken 1987]), la relation entre l'attitude et le comportement (exemple : [Fazio 1990]), la formation des impressions (exemple : [Brewer 1988]). Les travaux plus récents se sont employés à proposer des modèles généraux identifiant des principes psychologiques fondamentaux indépendamment du domaine. Ces modèles généraux peuvent être distingués selon deux finalités ; les modèles généraux des « Dual process theories » qui décrivent les processus mentaux en deux principes opératoires (exemples : [Epstein 1994] [Kahneman 2003]), et les modèles formels qui quantifient la contribution de processus mentaux distincts dans la réponse comportementale en termes de modélisations mathématiques (exemples : [Payne 2008] [Sherman 2008]).

Dans la modélisation de la « dual process theory », les processus automatiques sont définis comme tels au regard de 4 conditions : (1) ils sont provoqués de manière non intentionnelle, (2) ils requièrent peu de ressources cognitives, (3) ils ne peuvent pas être arrêtés volontairement et (4) ils sont mis en œuvre de manière inconsciente [Bargh 1994]. Les processus contrôlés sont quant à eux caractérisés comme : (1) initiés intentionnellement, (2) consommant une quantité importante de ressources cognitives, (3) pouvant être arrêtés volontairement et (4) opérant de manière consciente.

Le premier modèle général de « Dual Process Theory » usuellement cité comme référence initiale est celui de Epstein [Epstein 1973]. Nommé CEST, il décrit deux systèmes appelés le système expérientiel (the experiential system) et le système rationnel (the rational system). Le système « experiential » opère de manière automatique, sans effort et sur la base de connexions associatives fortement liées aux affects et aux notions de plaisir et de souffrance.

Les changements de ce système sont lents et sont la conséquence d'expériences répétées et relativement intenses. Les principes opératoires de ce système prennent leurs racines dans la structure du cerveau qui s'est développée en premier et qui n'a pas été remplacée par des évolutions plus récentes servant de fondation à la construction du second système : le système « rational ». Les processus de ce système « rational » sont caractérisés par un aspect intentionnel et des relations logiques entre les éléments qu'ils manipulent lors d'un effort conséquent. Une représentation de la réalité est encodée dans le système « rational » à travers des symboles abstraits, des mots et des nombres. Les changements de ce système sont plus rapides, comparés à ceux de « experiential system », et dépendent du poids des arguments logiques et de la disponibilité de nouvelles évidences. Pour Epstein les deux systèmes opèrent en parallèle de manière à ce que chacun puisse produire ses propres réponses comportementales. Dans le cas où ces réponses ne seraient pas alignées, l'individu tend à ressentir un conflit « entre la tête et le cœur » (« conflit between the head and the heart » p.710 [Epstein 1994]). En même temps, les deux systèmes peuvent interagir l'un avec l'autre de telle sorte que les processus inconscients influencent continuellement ceux du système conscient. Néanmoins, cette interaction entre les deux systèmes est décrite par Epstein comme asymétrique puisque les processus du « experiential system » sont inconscients et donc difficilement contrôlables par le « rational system ».

Le modèle CEST est un cadre conceptuel pour l'interprétation d'observations plus qu'il n'est une source de déduction de prédictions que l'on puisse tester. Néanmoins il est cité comme le premier modèle du genre et la plupart des hypothèses fondamentales qu'il intègre sont souvent retrouvées dans les propositions de ses successeurs.

Une autre étape importante du développement de la théorie générale du « Dual-Process » en psychologie cognitive est incarnée par les travaux de Smith and DeCoster's [Smith & DeCoster 2000]. Les auteurs argumentent que les phénomènes identifiés et étudiés dans le cadre des recherches en domaines spécifiques mettent en lumière les opérations de deux types de mémoire distinctes. Une mémoire associative, à l'apprentissage lent, et une mémoire basée sur des règles, à l'apprentissage rapide. Les processus de la mémoire associative sont spécifiés comme étant structurés par les similarités et la continuité d'associations simples entre objets et événements lentement appris au cours de l'expérience. Les processus de la mémoire basée sur les règles sont quant à eux structurés par le langage et la logique. Ils sont représentés par des règles utilisant des symboles et qui peuvent être apprises rapidement à partir d'une expérience restreinte.

Travaillant à l'intégration des théories précédentes pour des recherches sur les heuristiques et les biais cognitifs en contexte économique, Kahneman [Kahneman 2003] propose une théorie générale du « Dual-Process » qui reprend de nombreuses caractéristiques des modèles de Epstein et de Smith & DeCoster's. Il y distingue deux systèmes nommés « Système 1 » et « Système 2 » (une terminologie proposée par Richard West et Keith Stanovich [Stanovich & West 2000]) qui illustrent la dichotomie entre intuition (système 1) et raisonnement (système 2). Pour Kahneman, les processus du système 1 sont rapides, parallèles, automatiques, nécessitant peu d'effort, associatifs, à apprentissage lent, et basés sur les émotions. En contraste, les processus du système 2 sont lents, séquentiels, contrôlés, nécessitant beaucoup d'efforts, gouvernés par des règles, à apprentissage rapide, neutres émotionnellement. Les productions du système 1 sont considérées comme des impressions intuitives alors que celles du système 2 sont comme des jugements qui peuvent être basés sur les impressions intuitives ou sur un raisonnement délibéré. En ce sens, pour Kahneman, l'une des fonctions importantes du système 2 est de monitorer l'activité et les entrées du système 1.

Bien que Kahneman ait repris des hypothèses provenant des travaux de ses prédécesseurs, sa ré-analyse et ses résultats empiriques concernant les heuristiques et biais cognitifs en contexte de prise de décision économique ont grandement contribués au développement de la « Dual Process theory » et à sa démocratisation vers les autres disciplines scientifiques. En témoigne son invitation à participer à une discussion avec les récents lauréats du prix Turing, Geoffray Hinton, Yann LeCun et Yoshua Bengio, dans le cadre de l'organisation de la 34ème conférence en intelligence artificielle AAAI 2020.

En psychologie l'un des modèles les plus influents à l'heure actuelle, selon le rapport [Gawronski & Creighton 2013], est celui de Strack et Deutsch : RIM [Strack & Deutsh 2004]. RIM propose l'hypothèse que les comportements sociaux sont guidés par deux systèmes de traitement de l'information qui opèrent simultanément ; l' « impulsive system (IS) » et le « reflective system (RS) » qui, comme leurs noms l'indiquent, correspondent respectivement au système 1 et au système 2. Bien que les deux systèmes opèrent en parallèle l'un de l'autre, le modèle RIM met l'accent sur la priorité du système impulsif (S1). Ceci dut au fait que le système réflectif n'entre en action que dans la condition où les ressources cognitives le permettent, alors que le fonctionnement du système impulsif en est considéré indépendant. En détail, les caractéristiques des deux systèmes sont similaires à celles que nous avons présentées précédemment. Les spécificités du modèle sont qu'il intègre un rôle

important à la motivation, un processus du système impulsif, et intègre des principes basiques d'homéostasie et de régulation. L'utilisation de ce modèle a servi d'hypothèse à plusieurs expériences empiriques relatives aux comportements individuels comme par exemple pour le rôle des impulsions et du contrôle dans l'autorégulation comportemental [Hofmann et al. 2007] ou les conséquences émotionnelles de la privation de nourriture [Hoefl et al. 2009].

Un autre courant de modélisation de la « dual process theory » prend les traits de modèle formel, décrit en termes de formules mathématiques, dont l'objectif est de modéliser les procédures de la psychologie cognitives pour quantifier les contributions respectives des processus automatique et contrôlés dans la réponse comportementale. Des auteurs comme Jacoby ont conçu des modèles dans lesquels les processus contrôlés sont dominant [Jacoby 1991] et d'autres où ce sont les processus automatiques qui ont la priorité [Lindsay & Jacoby 1994]. L'hypothèse commune à ces modélisations est que l'une des deux classes de processus n'opère que dans le cas où les processus de l'autre classe échouent. Prenant le contre-pied de cette position dichotomique certains auteurs ont proposé des modèles plus complexes comme par exemple celui de [Conrey et al. 2005] qui articule non pas deux mais quatre systèmes. Néanmoins cette hypothèse « multi-process » garde la notion intrinsèque de distinction entre processus automatique et processus contrôlés. A l'extrême opposé, incarnant la principale critique de la « dual process theory », certaines propositions ne font état que d'une seule classe de processus du type « Si-Alors » en argumentant le fait que les processus automatiques peuvent aussi être conceptualisés de la sorte et que leurs différences avec les processus contrôlés sont dues au fait de variables internes spécifiques [Kruglanski 2006].

Ces considérations de « single process theory » et « multiple process theory » restent floues et partiellement explorées par la communauté qui débat actuellement de leurs pertinences conceptuelles. Une question de recherche pourrait ainsi être formulée comme « Combien de classe de processus il y a-t-il ? ». Factuellement, la grande majorité des travaux théoriques et des résultats empiriques vont dans le sens de la « dual process theory ».

Il nous semble essentiel d'ajouter en référence le caractère historique de la « dual process theory » qui a traversé les siècles depuis l'antiquité, déguisé derrière différents noms, vocabulaires et finalités. Comme nous l'avons écrit plus haut, la plupart des idées dépeintes dans le cadre psychologique de la « Dual Process Theory » constituaient déjà la plupart des

réflexions philosophiques des grands penseurs de notre histoire. On trouvera chez Epicure les prémisses de la distinction des désirs, chez Spinoza l'exploration des affects (S1) et de l'entendement (S2), chez Kant les limites de la raison (S2), chez Schopenhauer l'articulation particulière entre volontés et représentations (S1/S2), chez Nietzsche la puissance de la volonté (S1), chez Freud et Jung les abîmes de l'inconscient (S1).

Pour résumer, en 1890 le philosophe et psychologue William James propose de décomposer la cognition de l'homme en deux systèmes qu'il nomme « Associative thinking » et « Reasoning thinking » [James 1884]. De cette hypothèse naîtra un courant théorique, appelé « Dual Process Theory », pour lequel l'esprit humain résulte de la coopération de deux entités; celles-ci prendront différents noms selon les auteurs (Raisonnement inconscient / raisonnement conscient, involontaire / volontaire, implicite / explicite, automatique / contrôlé, rapide / lent, subjectif / objectif etc.). Dans la suite de ce rapport, nous nommons ces entités « Système 1 » et « Système 2 », un vocabulaire inventé par Richard West et Keith Stanovich [Stanovich & West 2000] et que nous adoptons pour la neutralité de sa sémantique. Dans son livre « Système 1 / Système 2 : Les deux vitesses de la pensée », le prix décerné Nobel d'économie Daniel Kahneman décrit le système 1 comme un mécanisme qui « fonctionne automatiquement et rapidement, avec peu ou pas d'effort et aucune sensation de contrôle délibéré » et le système 2 qui « accorde de l'attention aux activités mentales contraignantes qui le nécessitent, y compris des calculs complexes » [Kahneman 2011]. Pour Kahneman, l'individu associe son soi conscient, qui raisonne et possède des convictions, au système 2 qu'il qualifie de lent. À l'inverse, toujours selon Kahneman, le système 1 est rapide et produit sans effort des actions automatiques qui engendrent des enchaînements d'idées complexes. Dans les parties suivantes nous explorerons les concepts « d'intuition » et « d'émotions » qui sont associés au système 1, puis ceux de « raison » et de « logiques » associés au système 2.

2.3.3 Intuitions et émotions

Pour Daniel Kahneman, les intuitions sont des pensées et des préférences qui arrivent à l'esprit de manière quasi instantanée et sans trop de réflexion a priori [Kahneman 2003]. Selon lui, les jugements intuitifs occupent une position qui se trouve entre le traitement automatique des perceptions et les opérations délibérées du raisonnement. Le contenu de ces jugements intuitifs est déterminé par l'expérience et la personnalité de l'individu face au contexte. Dans plusieurs de ses travaux il ajoute que les affects, parfois nommées émotions,

sont des évaluations automatiques constituant le principal déterminant des jugements et comportements intuitifs [Kahneman 1994]. L'hypothèse majeure concernant le rôle de l'intuition dans le comportement global est qu'elle constitue la source de la plupart des jugements même quand ces derniers sont dominés par la raison [Haidt 2001]. Cela étant dû au fait que la raison utilise des représentations mentales dont la forme et le contenu sont subordonnés aux processus intuitifs qui les ont rendus accessibles à la conscience. David Myers réaffirme ce point et explique que les processus automatiques inconscients constituent la majeure partie de l'activité mentale d'un individu. Pour lui ces processus incluent les réactions émotionnelles instantanées, la mémoire implicite et les perceptions subliminales [Myers 2002]. Dans son article, Myers rapporte les résultats d'expériences marquantes ; ayant perdu une portion de leur cortex visuel en conséquence d'une attaque cérébrale ou d'une intervention chirurgicale, les individus observés sont aveugles d'une partie de leur champ de vision. Lorsqu'ils disposent des bâtonnets dans le champ visuel dysfonctionnel, les patients rapportent ne rien voir. Cependant lorsqu'on leur demande de deviner l'orientation verticale ou horizontale des bâtonnets, ces mêmes patients répondent correctement. Cette hypothèse très forte concernant l'implication des processus inconscients en aval des processus conscients fut formulée explicitement par Benjamin Libet dans un article de 1985 au titre évocateur : « Unconscious cerebral initiative and the rôle of conscious will in voluntary action » [Libet 1985]. De nombreux travaux sont venus la corroborer, l'un des derniers en date étant l'étude neurologique de Roger Koenig-Robert [Koening & Pearson 2019]. Cette place singulière qu'occupent les processus inconscients et l'intuition qui en découle a également été théorisée en philosophie de l'esprit par des auteurs de renom comme Baruch Spinoza [Spinoza 1677], Carl Jung [Jung 1964], Sigmund Freud[Freud et al. 1954] ou Jacques Lacan [Lacan 2013]. Chacun à sa manière, ils ont dépeint à travers leurs théories, l'existence d'une entité cognitive discernable de la pensée consciente. Une entité qui s'exprime à travers les émotions ou dans les rêves.

Dans ses « Lettres à une princesse allemande », Euler avait commencé à philosopher sur le compte de l'intuition [Euler 1775], lui donnant un rôle bien spécifique dans la cognition. Pour Arthur Schopenhauer la totalité du monde de l'expérience est conditionnée par les formes spéciales de l'intuition que sont le temps, l'espace et la loi de causalité ; formes discutées en premier lieu dans la théorie de l'esthétique transcendantale de Kant, puis développées par Schopenhauer. Pour ces auteurs ainsi que pour beaucoup d'autres de leurs successeurs, l'intuition est essentiellement intellectuelle et non simplement sensible, les sens ne donnent que la sensation qui n'est pas encore l'intuition. La sensation donne l'effet, et

l'intuition, comme fonction du cerveau, en cherche la cause (chapitre 2 du supplément « La théorie de la représentation intuitive » du livre 1 « Le monde comme représentation » d'Arthur Schopenhauer [Schopenhauer 1888]).

L'émotion est un concept flou et sa théorisation est encore sujette à discussion. Pour Spinoza un affect est un état de l'âme. Il s'agit d'une affection du corps en même temps que du mental. Les affects fondamentaux de Spinoza sont le désir, la joie et la tristesse. Si l'émotion est un état du corps et le sentiment (« feeling ») un état de la psyché, le concept d'affect est une combinaison des deux. En 1884 William James lance le débat [James 1884] en se demandant « Qu'est-ce qu'une émotion ? ». Pour lui, l'émotion est une production du cerveau qui prend conscience de processus viscéraux survenant dans le corps. En 1920, Walter Cannon fait remarquer que les réponses corporelles sont sous contrôle du système nerveux autonome. Pour lui, les émotions sont une production du cerveau mais ne nécessitent pas que ce dernier « interprète » quoi que ce soit en provenance du corps [Cannon 1920]. En 1960 la position évolue sous l'influence de Stanley Schachter et Jerome Singer et s'établit en un juste milieu : L'émotion surgit lorsque les capacités cognitives fournissent une explication à des signaux corporels ambigus. Cette notion d'explication est reformulée en termes d'évaluation par Magda Arnold [Arnold 1960] et, de la même manière, l'émotion est considérée comme un phénomène émergeant des suites de l'évaluation qu'effectue le cerveau d'une situation pour laquelle il est nécessaire de déterminer si elle est « bonne » ou « mauvaise ». Enfin Robert Zajonc finit par montrer que les émotions peuvent être indépendantes de la cognition et peuvent même exister avant toute activité cognitive, ouvrant la voie aux études sur la perception inconsciente et sur la recherche contemporaine sur les émotions qui considèrent que nos réactions émotives peuvent survenir en l'absence de la conscience explicite d'un stimulus [Zajonc 1998]. Dans la littérature moderne le rôle des émotions est souvent décrit en référençant la théorie des marqueurs somatiques de Antonio Damasio [Damasio 1996]. Il y suggère l'influence de « signaux » dans le traitement, inconscient ou conscient, de stimulus. Ces signaux appelés marqueurs somatiques proviennent de processus de bio régulation qui s'expriment sous la forme d'émotions (inconscient) ou de sentiments (conscient) mais pas uniquement. Ils font référence à des états du corps et de ses régulations, même quand ces états ne proviennent pas du corps en lui-même mais d'une représentation mentale de ce dernier. La théorie des marqueurs somatiques ne relate donc pas uniquement des émotions, c'est pour cela que Damasio utilise le terme « somatique ».

Nous n'irons pas plus en détail sur les tournants qui ont été pris dans ce domaine et tenterons plutôt de résumer les hypothèses qui semblent être communément acceptées par la communauté scientifique : (1) Les émotions sont une production du corps, mais l'activité du corps est elle-même en partie déterminée par les processus cognitifs. (2) La valence des émotions sert d'évaluation quant aux états du corps et des perceptions de l'environnement.

2.3.4 Raisonnement conceptuel et logique

Le dictionnaire Larousse en ligne définit le verbe raisonner comme l'action de faire usage de sa raison, de sa capacité de réflexion. Il s'agit de « lier logiquement entre elles des propositions pour aboutir à une proposition nouvelle, à une conclusion ». Pour le mot concept, ce même dictionnaire note : « Idée générale et abstraite que se fait l'esprit humain d'un objet de pensée concret ou abstrait, et qui lui permet de rattacher à ce même objet les diverses perceptions qu'il en a, et d'en organiser les connaissances ». Dans le langage courant le raisonnement conceptuel est donc l'acte de lier logiquement des objets de pensée pour aboutir à de nouveaux objets de pensée ou connaissances. La notion de « logiquement » relève de l'application de règles (logiques) qui déterminent la manière de lier les concepts entre eux et s'apparente à la syntaxe d'un langage dont les concepts sont les mots. On peut dire que la philosophie est l'art de raisonner en utilisant la langue naturelle, un langage informel, alors que les mathématiques sont l'art de raisonner en utilisant les nombres, un langage formel. Néanmoins dans les deux cas il s'agit bien de raisonnement conceptuel, les nombres ou les mots sont associés à des concepts liés en une phrase en suivant une logique syntaxique.

Des études philosophiques ont été menées de manière très approfondie par Emmanuel Kant qui publia en 1781 le livre « Critique de la raison pure » dans lequel il rationalise le raisonnement conceptuel par l'exploration de ses limites. Ces limites sont principalement les conséquences de la nature même des concepts qui, comme le stipule le dictionnaire Larousse, sont des objets de pensée concrets ou abstraits. Cela signifie qu'ils n'ont rien de réel et ne sont que des représentations, des modélisations imparfaites de la réalité formées à partir des perceptions et de l'expérience. En effet, les perceptions sont bornées par les capacités des organes senseurs qui les génèrent. Les sens n'ont pas accès à l'intégralité des phénomènes physiques réels, ils n'en mesurent qu'un certain spectre. Cela signifie que le raisonnement sur la réalité, quel que soit son support (philosophique, mathématique ou autre), se fait sur une base d'informations incomplètes. On retrouve là l'une des contraintes

impliquant la notion de rationalité limitée formulée par Herbert Simon [Simon 1972]. Pour Arthur Schopenhauer, les concepts forment une « classe spéciale de représentations, [...] distinctes des représentations intuitives [...] car elles n'existent que dans l'esprit humain » qu'il nomme « représentations de représentations ». Le philosophe Allemand procède à une analyse détaillée de ce rapport entre concepts et perceptions dans son livre « Le monde comme volonté et comme représentation » [Schopenhauer 1888] et plus particulièrement dans le chapitre 9. Son essai fait suite aux travaux de Kant et a été publié en 1819. Pour lui, et de manière cohérente avec ce qui a été écrit plus haut, le raisonnement conceptuel est une « reproduction du monde » qui repose sur l'intuition « d'où il tire son intelligibilité ». Il écrit :

« Le monde tout entier de la réflexion repose sur le monde de l'intuition et y a ses racines. L'extrême évidence, l'évidence originelle est une intuition, comme son nom même l'indique : ou bien elle est empirique, ou bien elle repose sur l'intuition a priori des conditions de la possibilité de l'expérience. [...] Tout concept n'existe et n'a de valeur qu'autant qu'il est en relation, aussi lointaine qu'on voudra, avec une représentation intuitive ; ce qui est vrai des concepts est vrai des jugements qu'ils ont servi à former, et aussi de toutes les sciences. »

Cette thèse reprend et complète des réflexions philosophiques nées dans l'antiquité, par exemple dans les pensées d'Epicure. Elle trouvera également écho dans la philosophie de William James qui, en 1907 dans son le livre « Le pragmatisme » [James & Burkhardt 1975], définit les concepts comme des constructions, des instruments par lesquels l'homme met son expérience en ordre pour mieux la maîtriser. Dans cette perspective la vérité est entièrement subjective et constitue la propriété attribuée aux idées qui se sont avérées fonctionnelles et utiles à notre appréhension du monde. Les concepts utilisés par l'Homme sont alors des constructions subjectives et ne sont en aucun cas universels ou a priori définis. Les conclusions des raisonnements qui s'en servent comme matière première n'en sont pas plus universelles. Ce sont également des représentations fondamentalement bornées par les capacités perceptives et l'expérience [James & Burkhardt 1975].

« Les sensations qu'il aura prises par le toucher seront, pour ainsi dire, le moule de toutes ses idées » (Denis Diderot, Lettre sur les aveugles, 1875)

Toujours d'après cette position, pour ce qui est des logiques utilisées par le raisonnement conceptuel Humain : étant donné que les concepts ne sont pas a priori définis, aucune logique particulière sur la manière de les associer ne peut être a priori définie. Seul existerait

un phénomène général qui est la logique associative et qui permet la liaison de deux concepts. Cette logique originelle est appliquée sur un critère de temporalité. Deux objets de pensée ayant une activation simultanée ou relativement proche dans le temps tendent à créer une association. La tendance humaine à définir une causalité entre deux événements corrélés temporellement en serait l'artefact le plus saillant. Les logiques suivantes, des associations particulières, seront acquises en tant que construction intuitive. Schopenhauer écrit à ce sujet :

« L'intuition [...] est la source de toute vérité et le fondement de toute science. Il faut en excepter seulement la logique, qui est fondée sur la connaissance non intuitive, quoique immédiate, qu'acquiert la raison de ses propres lois. »

Notons que nous utilisons ici le mot logique pour désigner les règles qui permettent de lier des concepts entre eux. Nous ne parlons pas de la logique mathématique qui a été construite par l'homme comme un outil formel. La logique dans le sens mathématique est un produit raffiné à partir de la matière brute qu'est le langage courant ; langage courant qui a lui émergé des interactions socio-culturelles des hommes pour servir d'outil de communication. Les chercheurs Aleksander et Burnett [Aleksander & Burnett 1984] commentent ce fait d'une manière qui fait écho aux propos de Schopenhauer rapportés ci-dessus :

« Il se pourrait que les arbres de décision, la logique du premier ordre, etc. soient seulement des rationalisations de faits de traitement de l'information qui sont, en réalité, réalisés par des moyens très différents. Une sorte de commentaire, - de vernis -, que le cerveau produit pour rendre compte de son propre comportement. »

Poincaré écrivait également :

« La logique formelle n'est autre chose que l'étude des propriétés communes à toute classification ; elle nous apprend que deux soldats qui font partie du même régiment appartiennent par cela à la même brigade, et par conséquent à la même division. » (Rapporté par [Meyerson 2011])

D'après les philosophes, les concepts et les logiques qui servent de matière première au mécanisme du raisonnement conceptuel sont des constructions subjectives basées sur l'expérience et l'intuition. De nombreuses observations empiriques corroborent cette hypothèse. Les logiques des langues naturelles en sont un bon exemple, elles possèdent des

alphabets, des mots et des syntaxes différents selon les cultures. Les concepts auxquels les mots du langage réfèrent ne sont pas non plus généralisables à travers les langages. Les règles qui subordonnent certains des comportements humains ne sont pas universelles mais résultent de constructions socio-culturelles, comme le montre par exemple une étude menée récemment sur la morale de personnes de cultures différentes [Awad et al. 2020].

2.4 Architecture cognitive

Cette partie présente les recherches bibliographiques menées sur la discipline de l'architecture cognitive, un sous-domaine de l'intelligence artificielle. L'objectif de cette partie est de discuter les hypothèses, connaissances et théorie qui ont été proposées dans diverses ressources que nous avons sélectionnées.

L'idée principale qui cherche à être transmise au lecteur pourrait synthétiquement se formuler comme suit: La théorie psychologique du «dual-Process» semble être sérieusement prise en considération par les acteurs de l'architecture cognitive. Il existe de nombreux modèles qui se distinguent par leur finalité. Il existe des architectures dont la finalité est de servir de modèle théorique pour expliquer des observations empiriques faites en neurosciences. Il existe d'autres architectures dont la raison est d'être implémentable informatiquement afin de tester des hypothèses ou de concevoir des agents autonomes.

2.4.1 L'architecture cognitive comme discipline

Une architecture cognitive décrit la structure et les processus globaux d'un modèle cognitif utilisé pour une analyse large, à plusieurs niveaux, de la cognition et du comportement [Sun 2004]. Elle spécifie les principes organisationnels fondamentaux d'un système complexe pour donner du sens aux relations entre fonctions et structures. Il s'agit, pour les acteurs de cette communauté, de prendre de la hauteur sur les diverses techniques de traitement de l'information et de les architecturer en un système global qui déploierait des capacités de niveau humain, ce que l'on appelle A.G.I « Artificial General Intelligence ». L'architecture cognitive est fortement liée au concept de système hybride, ces systèmes informatiques qui combinent les approches connexionnistes et symboliques. Néanmoins la conception de systèmes hybride dans le cadre de l'intelligence artificielle est une démarche d'ingénieur spécifique à un domaine, là où l'architecture cognitive se concentre sur l'accumulation de données empiriques pour produire des théories scientifiques qui les explique par des cycles

hypothèses – tests. Il faut donc distinguer deux approches dans l'architecture cognitive. L'approche d'ingénierie qui cherche à concevoir un système aux capacités domaine-génériques, une extension de la programmation de système hybride domaine-spécifique. Et l'approche scientifique où l'objectif est de proposer une modélisation de l'appareil cognitif humain qui sert à expliquer et comprendre les phénomènes psychologiques et les comportements.

Pour l'approche scientifique, l'architecture cognitive sert d'ensemble d'hypothèses qui peuvent être basées sur des observations empiriques en psychologie ou biologie, mais aussi sur des pensées et arguments philosophiques, ou encore sur des théories mathématiques relatives au calcul. Les disciplines mobilisées déterminent le niveau d'explication recherché, s'étalant du macroscopique au microscopique. Il y a par exemple le niveau sociologique où l'intention est d'expliquer les comportements collectifs ou interpersonnels, les mécanismes socio-culturels et organisationnels ou encore les interactions entre l'individu et son environnement. En dessous se trouve le niveau psychologique qui couvre les expériences individuelles comme les performances cognitives, la conceptualisation, les croyances et leurs dynamismes. Puis le niveau des composants ou l'attention est portée sur les éléments, ou groupes d'éléments, computationnels qui permettent d'opérer telle ou telle fonction cognitive. Enfin le niveau le plus microscopique est le niveau physiologique qui cherche à élucider le fonctionnement du substrat biologique de la cognition.

En 1994 Allen Newell, l'un des pionniers présents à la conférence de Dartmouth, publie un livre intitulé « Unified Theories of Cognition » [Newell 1994] dans lequel il argumente la nécessité de développer une théorie générale de la cognition humaine qui couvrirait l'ensemble des phénomènes cognitifs. Il s'agit d'une théorie qui unifierait les nombreuses « micro-théories » qui mettent en lumière différents mécanismes liés à la perception, au raisonnement, à l'apprentissage et au contrôle moteur chez l'Homme. Il y présente le modèle cognitif SOAR comme exemple d'une telle tentative mais précise que, pour lui, les sciences cognitives ne sont pas encore assez matures pour atteindre une unique théorie et qu'il est nécessaire qu'il y ait de nombreuses tentatives. À la suite de cette proposition initiale, de nombreuses architectures cognitives ont été proposées.

Chaque chercheur avance sa propre construction modélisatrice et il n'est pas facile de les énumérer exhaustivement tant elles diffèrent dans leurs natures, fonctions et niveaux de granularité. Sans une définition claire et générale de la théorie de la cognition, chaque

architecture est basée sur différents ensembles de prémisses et d'hypothèses qui rendent leurs comparaisons difficiles. Pour tenter de résoudre cette problématique, Newell avait décrit des critères fonctionnels [Newell 1980]. Ces critères imposent la nécessité de rendre compte de comportements flexibles, en temps réel, rationnels, d'une large base de connaissances, d'un apprentissage, de capacités linguistiques et d'une conscience de soi. Sur la même intention, Ron Sun définit des « désidératas » pour l'architecture cognitive en incluant des critères de réalisme par rapport aux aspects écologiques et biologiques de la cognition humaine [Sun 2004]. Il reste qu'une grille de lecture unique ne permet pas d'englober l'ensemble des propositions d'architecture qui sont conçues dans des contextes et pour des finalités différentes. Certaines sont essentiellement philosophiques comme la « Society of Mind » de Marvin Minsky [Minsky 1986]. D'autres cherchent à formaliser des théories produites à partir d'observation en neurobiologie comme la « Global Workspace theory » de Stanislas Dehaene [Baars 2005] à l'opposé des intentions d'ingénieurs qui tentent quant à eux de produire des architectures implémentables dans des machines pour réaliser diverses tâches robotiques parfois très spécifiques. A noter que les connaissances sur le fonctionnement du cerveau, les techniques algorithmiques et le matériel informatique sont en perpétuelle évolution, la conception qu'ont les Hommes de la cognition et les moyens techniques par lesquels ils cherchent à la simuler changent de décennies en décennies. En conséquence, la production de nouvelles propositions architecturales, plus à jour avec les connaissances et moyens modernes, forme un flux quasi continu.

Dans un rapport datant de 2016, les chercheurs J. K. Tsotsos et I. Kotseruba font état de pas moins de 300 propositions d'architectures [Kotseruba & Tsotos 2016-1][Kotseruba & Tsotos 2016-2]. Parmi elles, ils en ont sélectionné 84 qu'ils ont passées en revue et comparées. La comparaison porte sur les capacités des systèmes, la nature des traitements mis en œuvre et les applications pour lesquelles elles ont été conçues. De cette étude, Tsotos et Kotseruba identifient 8 critères autour desquels les architectures se distinguent les unes des autres : il s'agit des fonctions de perception, d'attention, de sélection d'action, de mémoire, d'apprentissage, de raisonnement, de métacognition et des applications pratiques. Nous avons regroupé ces éléments et leurs principales modalités dans la figure suivante:

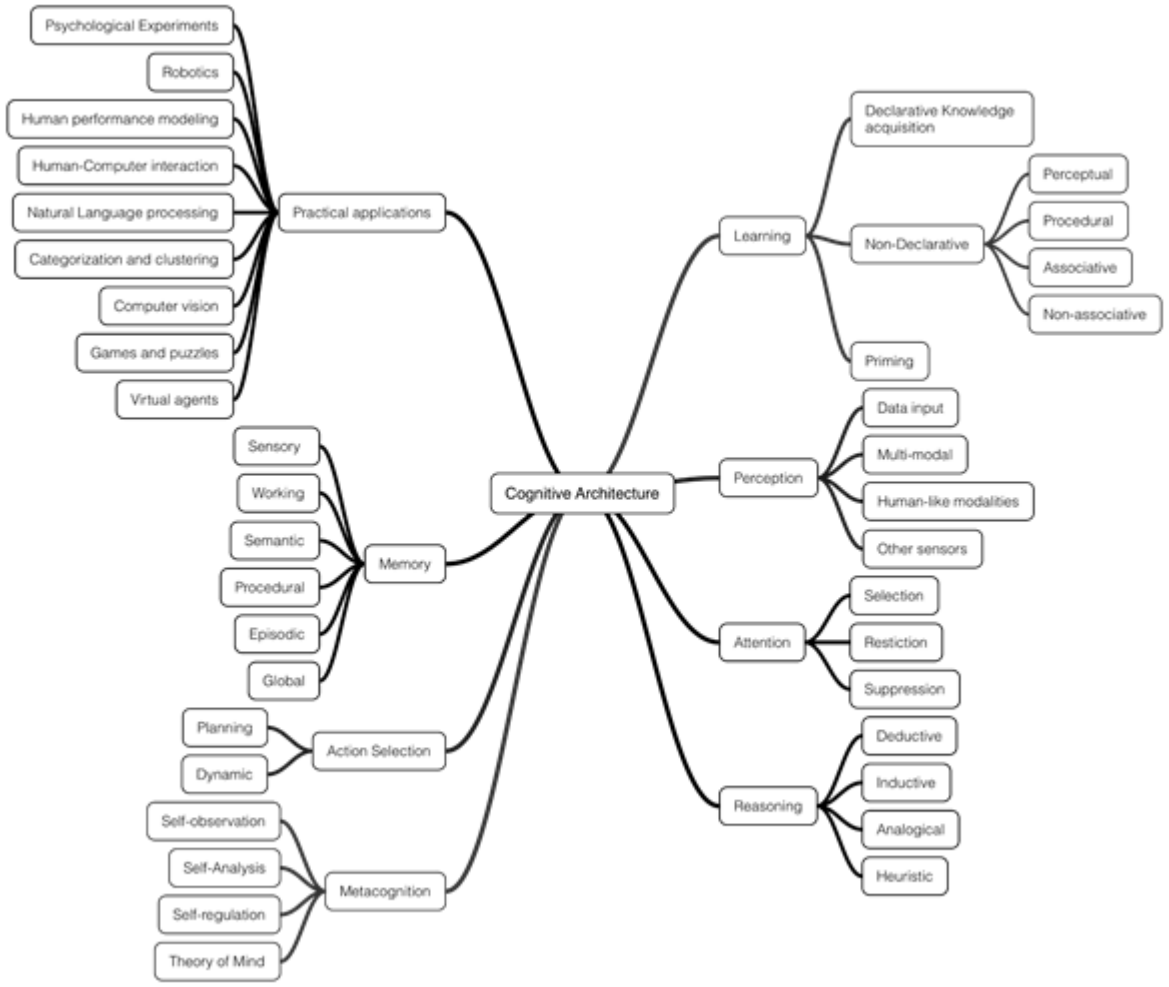

Figure 2.4.1 : *Critères de distinction entre architecture cognitive.*

2.4.2 Sélection d'exemples d'architectures cognitives

Nous proposons de parler plus spécifiquement de quatre modèles. Nous commençons par évoquer l'architecture ACT-R [Anderson et al. 2004], l'une des propositions les plus anciennes et figurant parmi les plus référencées du domaine. Ensuite nous présentons l'architecture CLARION [Sun 2007] qui se différencie notablement du modèle ACT-R. Ensuite nous présentons le modèle de Rasmussen [Rasmussen 1983] qui n'est pas à proprement parler considéré comme une architecture cognitive par la communauté de l'intelligence artificielle. Pourtant ce modèle a vocation à expliquer le comportement décisionnel de l'Homme et est très utilisé par les cogniticiens, il nous semblait donc intéressant de l'incorporer à notre discussion. Enfin nous présentons l'architecture CMC [Stocco et al. 2019][Laird et al 2017] qui se positionne à un haut niveau de généralité et

prétend à englober les caractéristiques fondamentales qui se retrouvent dans toutes les architectures cognitives.

2.4.2.1 ACT-R

L'architecture ACT-R [Anderson et al. 2004] est composée de modules qui sont le plus possible associés à différentes régions corticales du cerveau (Figure 2.4.2.1). Les « Perceptual-motor modules » sont responsables des actions de perceptions et de contrôle moteur. Le « Goal module » représente et garde en mémoire les objectifs et intentions du système. Le « Declarative module » récupère l'information depuis la mémoire pour son traitement. Les connaissances déclaratives sont intégrées au système par son concepteur et correspondent à des connaissances métiers à la manière d'un système expert. Chaque module place des morceaux d'informations dans une mémoire tampon où ils peuvent être détectés par le « production system » qui orchestre et coordonne l'activité des modules. Le « production system » répond en fonction des motifs d'informations présents dans les mémoires tampons. A chaque instant, une seule règle de production est sélectionnée pour être appliquée aux motifs détectés. Des processus sub-symboliques servent à guider la sélection des règles à utiliser. La sélection des règles est faite pour maximiser l'utilité du système au regard des objectifs et intentions définis dans le « goal module ». Cette sélection se fait sur une estimation de probabilité que les auteurs associent d'une certaine manière à un calcul bayésien. L'apprentissage consiste en un réglage des processus sub-symboliques. La théorie de l'architecture ACT-R ne définit pas précisément un nombre de modules fixe, leur nombre dépend de l'implémentation et des fonctions spécifiques à atteindre. Par exemple, dans la figure ci-dessous, un module « Visual module » appartenant à la famille des « Perceptual-motor modules » est responsable de l'identification d'objet dans le champ visuel et un « Manual module » a pour fonction de contrôler une paire de mains artificielles.

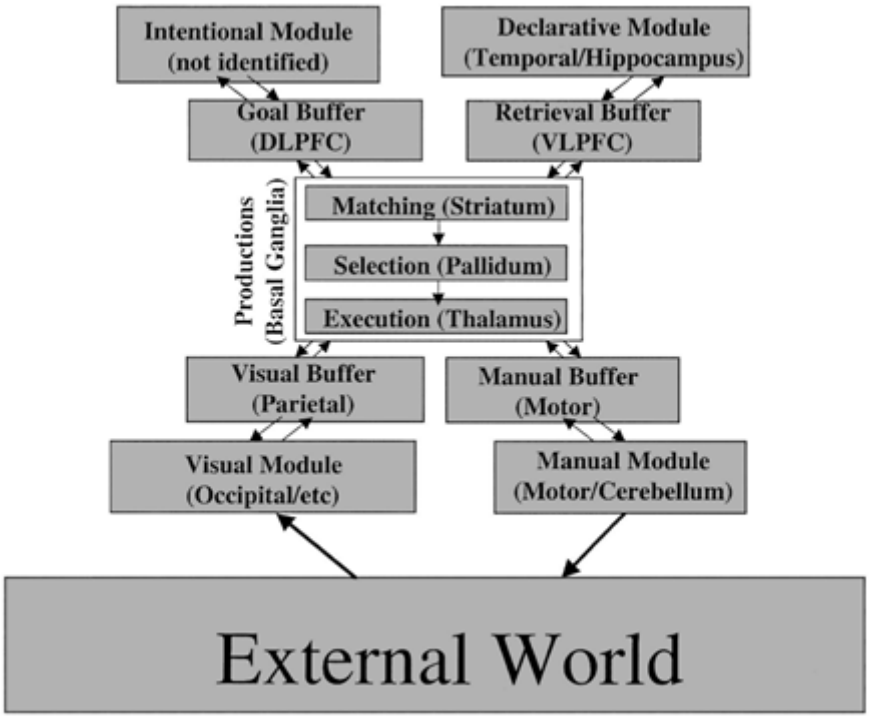

Figure 2.4.2.1 : *Architecture ACT-R.*

L'architecture cognitive ACT-R est un système modulaire avec un orchestrateur central (le carré «Matching», «Selection», «Execution»). Il utilise des connaissances déclaratives fournies par l'Humain lors de la conception afin de mettre en application des règles sélectionnées en fonction de l'estimation de probabilité de leur utilité vis-à-vis des objectifs eux aussi définis lors de la conception. Nous avons choisi de présenter succinctement ce modèle pour mettre en lumière deux notions courantes dans l'architecture cognitive que nous souhaitons discuter. Il s'agit de l'organisation modulaire et du caractère inné des processus déclaratifs.

L'organisation modulaire de ce type d'architecture est prototypique et se retrouve dans bon nombre de propositions. Sûrement inspirée par la théorie « Society of Mind » de Minsky et par la phrénologie qui cherchait à localiser différentes fonctions cérébrales dans le cerveau, elle ne saurait être considérée autrement que comme une hypothèse architecturale. Ce point de vue présente un intérêt certain pour l'implémentation de l'architecture en moyens informatiques, cependant comme le rapporte le psychiatre et neurologue Vinod Menon, la communauté neurologique se divisent actuellement sur ce point et l'hypothèse d'une architecture globale du cerveau en réseaux intrinsèquement entrelacés prend de plus en plus de poids au regard des phénomènes psychologiques et comportementaux qu'elle a

récemment pu expliquer [Menon 2011]. La deuxième particularité de l'architecture ACT-R se trouve au niveau de la spécification des traitements déclaratifs. Comme nous l'avons présenté ci-dessus, le modèle ACT-R utilise des règles et des objectifs qui sont définis en tant que connaissances déclaratives du système. Ces définitions sont « données » par l'Humain lors de la conception ou comme apport de connaissances métiers. Cela a également un intérêt en termes d'implémentation informatique mais l'hypothèse selon laquelle le cerveau humain serait initialement pourvu de telles connaissances est loin de faire l'unanimité.

Les notions d'orchestrateur symbolique central et de règle unique de ACT-R ne sont pas intégrées par toutes les architectures. Certaines rejettent ces hypothèses comme par exemple l'architecture de R. Sun nommée CLARION.

2.4.2.2 CLARION

Le modèle CLARION est constitué de 4 sous-systèmes distincts. Ces sous-systèmes sont nommés ACS pour « Action Centered Subsystem », NACS pour « Non-action Centered Subsystem », MS pour « Motivational Subsystem » et MCS pour « Meta-cognitive Subsystem ». Le rôle de l'ACS est de contrôler les actions, qu'elles aient pour but l'activation de mouvements physiques vers l'extérieur ou pour l'opération de processus mentaux interne. Le sous-système NACS quant à lui a pour objectif de maintenir des connaissances générales implicites ou explicites. Le « Motivational Subsystem », comme son nom l'indique, pourvoit l'état des motivations du système en termes d'intentions et de récompenses ce qui signifie qu'il indique si les résultats des traitements de perceptions, d'actions ou de cognition sont satisfaisants ou non. Enfin le rôle du MCS est de monitorer, diriger et modifier dynamiquement les opérations des autres sous-systèmes. Chaque sous-système possède deux niveaux de représentation ; implicite et explicite. Lors des opérations et des apprentissages, les deux niveaux coopèrent selon des approches « Bottom-up » et « top-down » intégrant ainsi, d'après Ron Sun, la théorie du « Dual Process ». L'approche « bottom-up » consiste en l'extraction de règles explicites à partir des actions implicites qui ont été jugées satisfaisantes. L'approche « top-down » quant à elle réfère à l'assimilation de routine implicite aux structures explicites. C'est ce que Sun appelle le « top-down » learning [Sun 2003].

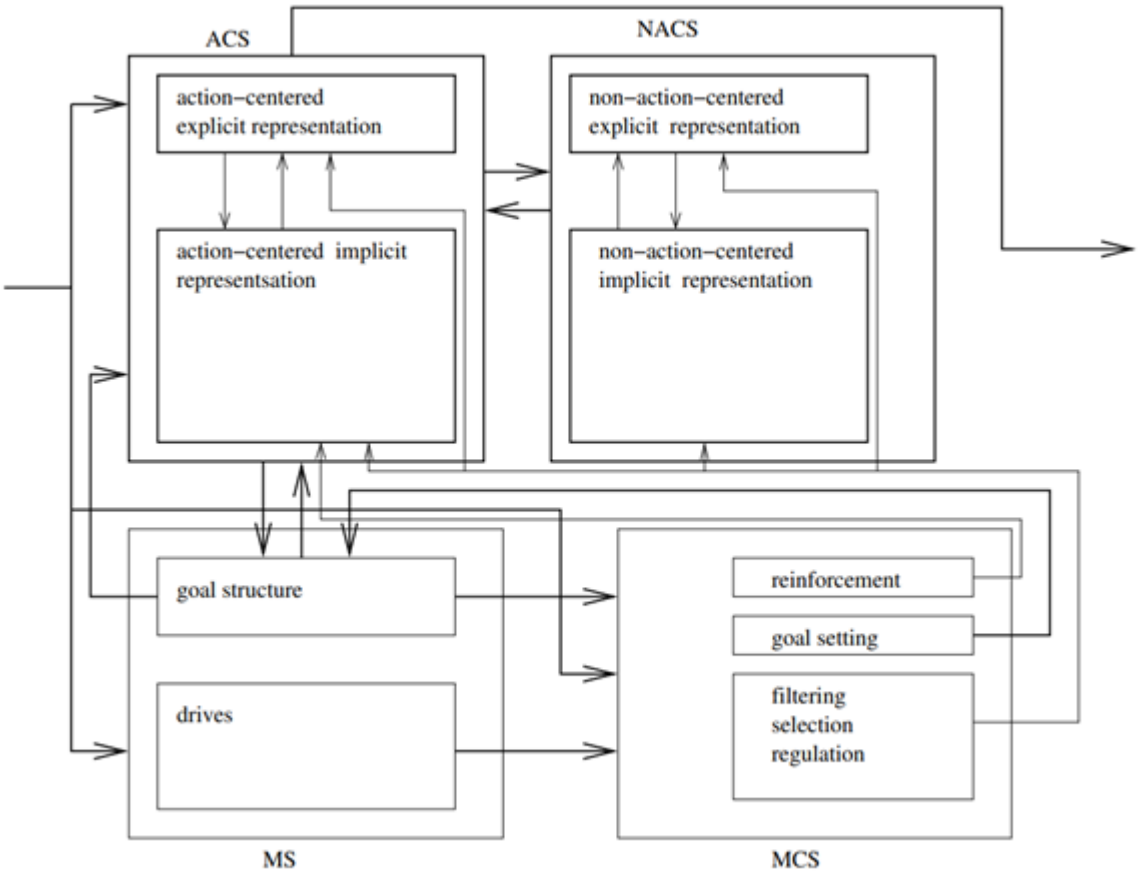

Figure 2.4.2.2 : *Architecture CLARION.*

L'architecture CLARION met en avant la décomposition en deux niveaux de représentation, implicite et explicite, et son auteur considère qu'elle intègre ainsi la Dual Process Theory ce qui n'était pas le cas d'ACT-R. Une autre architecture qui aborde cette notion de double forme de représentation est celle de Rasmussen.

2.4.2.3 Le modèle Rasmussen

Le modèle de Rasmussen [Rasmussen 1983] est conçu dans l'optique de fournir des prédictions relatives aux performances de l'Homme en interaction avec un système automatisé. (cf figure 2.4.2.3) L'auteur décompose le comportement humain en trois niveaux de performance : Le « skill based performance » (niveau des compétences), le « rule based performance » (niveau des règles ou procédures) et le « knowledge based performance » (niveau des connaissances). Les comportements « skill-based » représentent des performances sensorimotrices durant l'acte. Ils suivent une intention et sont mis en œuvre sans contrôle conscient en tant que motifs comportementaux automatiques et

fluides. Ces comportements utilisent une représentation implicite de l'environnement (« world-model »). Ils peuvent être modulés par des intentions conscientes. Les performances du second niveau de comportements, les « rule based behaviors », sont constituées de séquences de sous-routines mises en œuvre dans un environnement familier et contrôlées par des règles ou procédures. Ces règles peuvent avoir été dérivées à partir d'expériences antérieures, communiquées par un autre individu en tant qu'instructions ou « recette », ou avoir été préalablement préparées pour l'occasion comme solutions d'un processus conscient de résolution de problème ou de planification. L'élément particulier de ces types de comportement est qu'ils sont dirigés vers l'accomplissement d'un objectif et d'une procédure conscientisés alors que les « skill based » étaient des automatismes rendus ou devenus inconscients. Néanmoins l'auteur précise que la frontière entre les comportements « skill based » et « rule based » est floue car les actions misent en œuvre pour suivre les règles sont des séquences de sous-routines sensorimotrices. Face à des situations inconnues pour lesquelles l'individu ne possède pas de « recette » préétablie, le contrôle du comportement monte au troisième niveau de performance, appelé « knowledge based ». A ce niveau, l'objectif est explicitement formulé à partir d'une analyse de l'environnement et des intentions. S'ensuit une délibération qui considère plusieurs plans d'action qui sont testés physiquement dans un processus essai/échec (« trial and error ») ou conceptuellement en tant que simulation pensée des effets du plan sur l'environnement. A ce niveau de raisonnement les connaissances mobilisées sont représentées par une modélisation mentale explicite du monde. Rasmussen accompagne sa caractérisation des niveaux de performances individuelles par une catégorisation des formats de l'information utilisée. Pour lui, lors des comportements « skill-based » l'information est utilisée en tant que signaux espace-temps (« time-space signals ») représentant continuellement et quantitativement le comportement temporel et spatial de l'environnement. Il note que ces signaux n'ont pas de signification (« meaning ») autre que de l'information physique brute. Au niveau supérieur, les comportements « rule-based » emploient des signes qui servent à activer ou à modifier des actions prédéterminées. Les signes font référence à des situations conventionnelles ou à des expériences antérieures, ils ne sont pas assimilés à des concepts ou à des propriétés fonctionnelles de l'environnement. Ils sont uniquement utilisés pour sélectionner les règles qui contrôlent les séquences de sous routines des comportements « skill based » et ne peuvent pas être employés au niveau du raisonnement fonctionnel pour créer de nouvelles règles ou prédire l'évolution de l'environnement. Enfin au troisième niveau l'information prend la forme de symboles faisant référence à des

concepts et à des propriétés fonctionnelles. Alors que les signes ont des références extérieures en termes d'états et d'actions relatifs à l'environnement, les symboles sont définis comme des représentations conceptuelles internes.

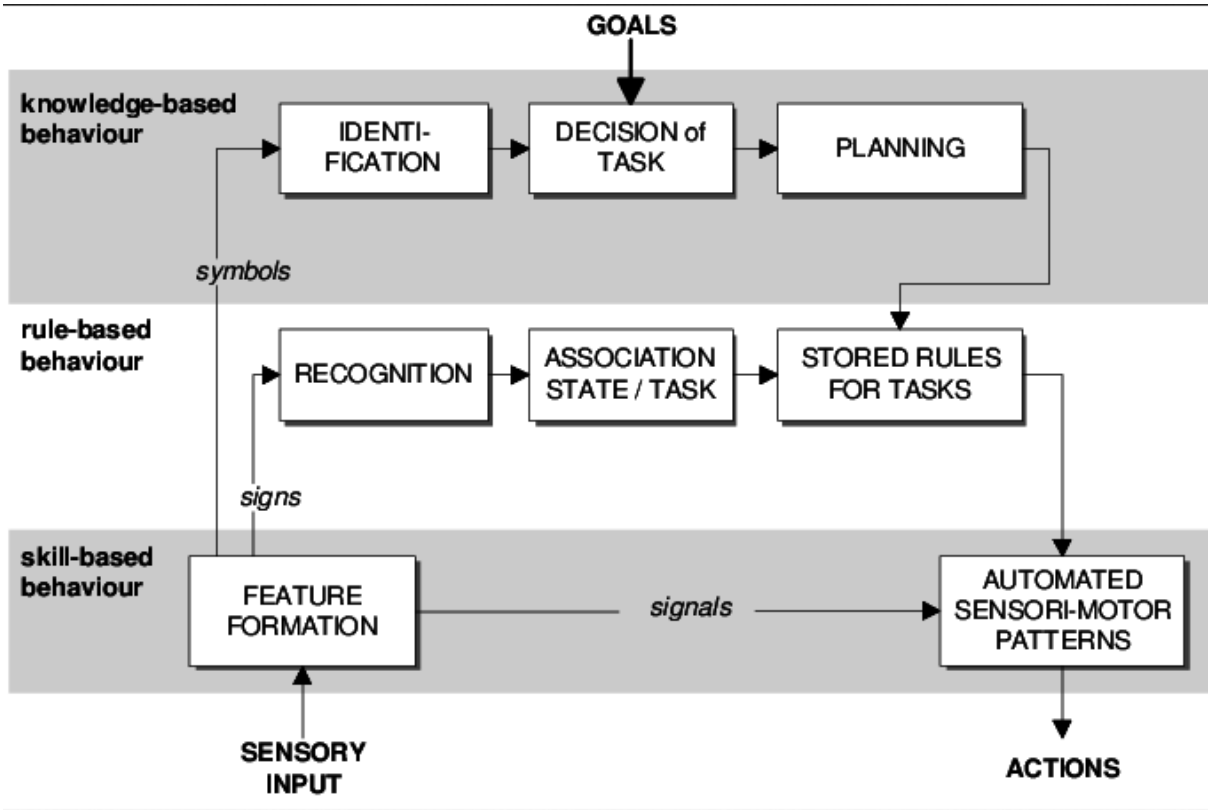

Figure 2.4.2.3 : *Architecture RASMUSSEN.*

Cette appréhension des comportements de l'individu humain en trois niveaux est courante et semble être une théorie à laquelle adhèrent de nombreux scientifiques de diverses disciplines. Nous proposons comme origine de ce point de vue l'essai de Alfred North Whitehead [Whitehead 1962], qui distingue trois catégories de performances humaines ; « instinctive actions », « reflex action » et « symbolically conditioned action ». Whitehead est co-auteur du « Principia Mathematica », avec Bertrand Russel que nous avions présenté dans la partie sur l'histoire du domaine de l'IA comme le maître à penser de la génération des pionniers de 1950.

L'architecture de Rasmussen se distingue de ACT-R et de CLARION, néanmoins on trouve des caractéristiques communes aux trois modèles. A vrai dire ces éléments communs sont partagés par la grande majorité des architectures. Observant ce constat, une architecture cognitive a été proposée dans le but de se positionner au niveau le plus général

possible pour encapsuler de manière minimal ces caractéristiques communes à toute architecture.

2.4.2.4 The standard Model of the Mind

Le modèle standard de l'esprit (« Standard Model of the Mind ») [Laird et al 2017], renommé « The Common Model of Cognition (CMC) » dans [Stocco et al. 2019], se présente comme un consensus sur l'architecture cognitive de l'homme et des systèmes artificiels de niveau humain. Il se positionne à un niveau de généralité qui lui permet, selon ses auteurs, de candidater en position de modèle à grande échelle du fonctionnement du cerveau humain. A cet effet les performances d'une de ses implémentations ont été comparées avec celles de quatre alternatives sur un critère de correspondance avec des données d'imagerie cérébrale de 200 participants humains réalisant un ensemble de 7 tâches sélectionnées pour couvrir l'ensemble des domaines d'activité cognitive. Les résultats rapportés par les auteurs montrent que leur modèle atteint une meilleure correspondance, entre performance du modèle et données neurologiques, que ses concurrents. Les auteurs prétendent donc proposer une architecture cognitive capable de modéliser l'intelligence humaine et artificielle tout en fournissant la meilleure explication des données neurologiques.

Le modèle présume que les agents déployant une intelligence comparable à celle de l'homme partagent cinq composants fonctionnels : Une mémoire déclarative à long terme, un ensemble de motifs état-actions provenant d'un apprentissage par renforcement et représentés dans une mémoire procédurale, une mémoire tampon correspondant à la mémoire de travail (court terme), un système de perception et un système d'action. La mémoire de travail est le point central (« hub ») par lequel tous les autres composants communiquent à l'exception des systèmes de perceptions et d'action qui possèdent une liaison directe. Le modèle inclut également un ensemble de contraintes portant sur le fonctionnement des mécanismes et des représentations qui caractérisent chaque composant. Pour les auteurs cette architecture encapsule les hypothèses et leçons apprises depuis les 50 dernières années en architecture cognitive. Toujours selon leurs dires, ces hypothèses se retrouvent dans toutes les propositions architecturales indépendamment de leur domaine d'application, que ce soit pour la conception d'agents artificiels ou pour la modélisation cognitive servant aux explications psychologiques et comportementales.

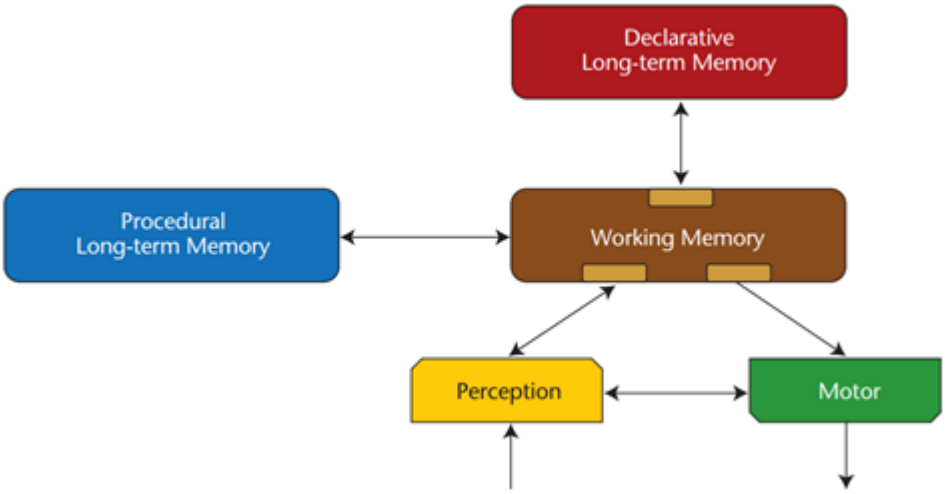

Figure 2.4.2-4 : *Architecture STANDARD MODEL OF THE MIND.*

De notre point de vue, la proposition des auteurs porte l'intérêt de mettre en lumière des notions qui semblent fondamentales pour une architecture cognitive et que l'on retrouve effectivement dans toutes les propositions ayant pour objectif de se rapprocher au mieux de la cognition humaine. La première notion est la dichotomie entre processus implicites et explicites, procéduraux et déclaratifs. Il s'agit de l'intégration de la « Dual Process Theory » présentée dans la partie sur la cognition humaine. On retrouve les prémisses de cette notion dans la philosophie antique d'Epicure et, de notre avis et relativement à nos connaissances, a été le plus précisément présenté dans les travaux philosophiques de Arthur Schopenhauer en 1819. Empiriquement, les travaux du psychologue Daniel Kahneman sont les plus représentatifs de cet état dualiste de la cognition. La deuxième notion est relative aux séquences des mécanismes inhérents à tout traitement de l'information. Conceptuellement, pour opérer un traitement d'information en utilisant des connaissances il est nécessaire que l'information soit intégrée (perception), modifiée (instructions contenues dans les mémoires) et utilisée (action).

2.4.3 Synthèse de l'étude des exemples d'architectures cognitives

Nous proposons de rassembler l'ensemble des hypothèses des architectures cognitives sélectionnées comme exemples, en deux principes :

1. La cognition est scindée en deux niveaux : implicite et explicite (Dual-Process Theory).

2. La cognition est un système de traitement de l'information (Perception / Traitement / Action).

On retrouve alors la proposition du CMC qui utilise un composant « Working memory » pour faire la jonction entre les deux modes de raisonnement et les organes sensori-moteurs. De notre point de vue, si le modèle CMC encapsule effectivement la majorité des architectures cognitives, c'est justement parce qu'elle est une représentation minimale de ces deux principes que toutes les architectures cognitives présupposent de manière plus ou moins explicite.

A la suite de cette recherche bibliographique nous avons exprimé un positionnement et une direction de recherche [Fruchart & LeBlanc 2019]. Nous y présentons une vision d'une architecture cognitive qui intègre la *Dual-Process Theory* et en amorçons une discussion sur l'intérêt de combiner les approches connexionnistes et symboliques dans une même architecture alors qualifiée de « hybride ».

Chapitre 3 : Modélisation I.P.S.E.L.

3.1 Démarche et hypothèses de modélisation

L'intitulé de notre sujet de recherche est : « Commandement et conduite des opérations aérospatiales. Aides à la décision et compréhension de situations complexes : confiance dans les données issues de systèmes utilisant l'intelligence artificielle ».

Les processus du commandement et de la conduite des opérations aérospatiales (C2Air) nous ont été proposés comme cadre applicatif. Le C2Air fait intervenir de nombreux acteurs humains et une multitude de machines. Ce domaine illustre bien une problématique moderne qui concerne de nombreuses activités : Comment assurer la bonne coopération entre l'Homme et des systèmes artificiels de plus en plus intelligents ?

Pour avancer dans ce sens, nous élargissons le périmètre de notre étude. Nous nous intéressons, de manière générale, aux interactions entre l'Homme et les systèmes utilisant l'intelligence artificielle (IA.), en apportant une attention particulière à la notion de confiance entre l'humain et la machine.

Le point de départ du travail a été d'interroger la nature et les finalités de la confiance de l'Homme envers un outil technologique. Les éléments de réponse que nous avons trouvés sont présentés dans le chapitre 2, à la partie 2.1 sur la confiance Homme/Automation. Nous les rapportons ici comme suit :

- La confiance est un état psycho-physiologique qui détermine la propension de l'opérateur à utiliser l'outil de manière appropriée.
- Chez l'individu, l'état de confiance est généré par sa cognition qui cherche à évaluer la possibilité que l'outil puisse l'aider à accomplir son objectif dans un contexte caractérisé par le risque et l'incertitude.

Un premier objectif serait de disposer d'un modèle de la cognition humaine pour pouvoir qualifier et analyser les mécanismes qui permettent la génération et la calibration de la confiance.

➔ Objectif 1: Disposer d'un modèle de la cognition humaine.

Pour présumer de la confiance attribuable aux systèmes utilisant l'IA, il nous faut également comprendre ces systèmes. Dans cette optique, nous avons exploré le domaine de l'intelligence artificielle avec une volonté d'acquérir une compréhension étendue et transversale de son contenu (Chapitre 2, partie 2.2, Le domaine de l'intelligence artificielle). Nous synthétisons les conclusions de ce travail bibliographique comme suit :

- L'IA est un domaine de recherche qui comprend différentes disciplines, courants et objectifs.
- Les systèmes utilisant l'IA sont des programmes informatiques qui utilisent une ou plusieurs techniques algorithmiques développées dans le cadre des recherches en IA.
- Le domaine est en constante évolution dans son périmètre, dans la définition de ses composants et dans les techniques employées.

Les systèmes utilisant l'IA sont donc multiples et les techniques qu'ils utilisent sont en perpétuelle évolution. Néanmoins ces techniques partagent certaines caractéristiques communes (Chapitre 2, partie 2.2.3, Approche connexionniste et approche symbolique). Pour pouvoir appréhender les interactions entre l'Homme et les systèmes utilisant l'IA, il nous semble alors nécessaire de diviser le travail en étudiant non pas l'IA de manière générale, mais des catégories de systèmes utilisant l'IA. Cette volonté nous amène à formuler un second objectif.

➔ Objectif 2: Disposer d'une catégorisation des systèmes d'IA.

Compte tenu de la rapidité d'évolution du domaine de l'IA, il est quasiment impossible de prédire ce que seront ses techniques, et les systèmes qu'elles vont engendrer, dans 5, 10 ou 20 ans. Cependant nous pouvons avancer que les systèmes utilisant l'IA auront toujours pour finalité d'imiter toutes ou parties des capacités cognitives humaines. Nous en déduisons qu'un modèle général des fonctions cognitives de l'Homme pourrait servir de référence (dans une optique du «*qui peut le plus peut le moins*») afin de classer les systèmes utilisant l'IA par rapport aux fonctions du modèle biologique qu'ils cherchent à imiter.

Cette réflexion nous laisse entrevoir une possibilité de regrouper les deux objectifs précédemment présentés en un seul et même objectif :

➔ Objectif final : Établir un modèle général de la cognition humaine exprimé dans le formalisme des systèmes de traitement de l'information.

Ce modèle aura deux utilités : (1) Servir de modèle comportemental pour considérer la cognition humaine et ses productions comme par exemple l'état de confiance, (2) Servir de modèle fonctionnel de référence pour catégoriser les systèmes utilisant l'IA.

En reprenant les résultats de la phase de recherche bibliographique sur la cognition humaine (Chapitre 2, partie 2.3, la cognition humaine) :

- La cognition humaine est une fonction complexe qui n'est pas parfaitement comprise. Toute modélisation sera nécessairement imparfaite.
- Il existe des modèles de la cognition humaine provenant des sciences économiques et sociales, de la psychologie, de la neurobiologie et de l'intelligence artificielle.
- De façon commune à ces modèles, la cognition semble se manifester à travers les phénomènes et notions relatives aux perceptions, aux émotions, à la gestion des connaissances, aux jugements, aux raisonnements et à l'apprentissage.
- Les modèles se différencient par les hypothèses qu'ils choisissent d'intégrer et par leur finalité. On distingue deux groupes de modèles. Il y a les modèles dont l'objectif est de rendre compte de phénomènes psychologiques, neurologiques ou comportementaux. Et il y a les modèles qui servent d'architecture pour la conception de systèmes informatiques et robotiques.

Dans l'état actuel de nos connaissances, nous ne trouvons pas de modèle répondant à toutes nos attentes et qui soit précisément adapté à nos besoins. A savoir : un modèle fonctionnel exprimé dans le formalisme des systèmes de traitement de l'information qui soit suffisamment complet pour décrire les productions de la cognition humaine et suffisamment général pour classifier les systèmes utilisant l'IA indépendamment de leurs implémentations techniques.

Les différentes lectures que nous avons menées durant notre recherche bibliographique nous ont permis d'explorer plusieurs disciplines et d'y entrevoir de nombreux ponts. Cette approche transdisciplinaire fut l'occasion de comparer et de compiler des connaissances et hypothèses hétérogènes qui portent néanmoins sur le même sujet : les mécanismes cognitifs de l'Homme. En l'occurrence nous avons l'impression que les contraintes qu'impose la finalité des modèles proposés (rendre compte de phénomènes empiriquement mesurables ou servir d'architecture implémentable) ont amené ces derniers à ignorer certains aspects jugés comme trop complexes, trop inconnus ou trop hypothétiques. Il s'agit par exemple des éléments concernant les mécanismes inconscients de la psyché, les émotions ou encore les

rapports entre intuition et rationalité. Cependant, ces aspects sont souvent discutés en philosophie de l'esprit. De plus, nous trouvons fertile de reprendre des théories explorées de longues dates comme celles d'illustres auteurs tel que Baruch Spinoza, Arthur Schopenhauer ou Carl Jung et de les reconsidérer à la fois dans un vocabulaire contemporain et au regard des observations empiriques récentes qui ont pu être faites grâces aux progrès de l'imagerie cérébrale et des capacités de simulation informatique. Nous décidons alors de saisir cette opportunité et de nous lancer dans la conception de notre propre modèle en nous basant principalement sur les connaissances théoriques en psychologie et philosophie de l'esprit, exprimées dans le formalisme moderne des systèmes de traitement de l'information, et guidées par une volonté de cohérence avec les expériences empiriques menées en intelligence artificielle sur les mécanismes d'apprentissages.

Toute tentative de modélisation passe par l'énonciation d'hypothèses prises comme fondations de l'objet à modéliser. Dans notre cas, nous voulons que notre modèle puisse rendre compte et expliquer un ensemble d'hypothèses que nous allons présenter ici et qui nous ont été inspirées des diverses études bibliographiques que nous avons pu faire. Pour indication nous y apposons les références aux œuvres qui nous ont inspirés. Cependant il n'est pas tout à fait juste d'attribuer une hypothèse à une référence particulière. Chaque auteur a contribué à sa manière et avec son vocabulaire à dépeindre un ou plusieurs aspects de la vie psychique de l'individu. Ces auteurs se sont influencés les uns les autres et il est difficile d'attribuer une idée à l'un plus qu'à l'autre. Au contraire, ces idées se sont raffinées à mesure qu'elles ont été reprises et développées par chacun de ces penseurs. La formulation que nous en faisons maintenant se nourrit de l'ensemble de ces travaux et les synthèses que nous en proposons doivent être prises comme des interprétations qui n'engagent que nous.

Liste des hypothèses que nous prenons concernant la cognition humaine :

Hypothèse 1 : Premièrement, nous suivons l'hypothèse selon laquelle la cognition humaine peut être modélisée comme la coopération de deux sous-systèmes nommés système 1 et système 2. Ces deux entités utilisent des connaissances distinctes que l'on peut qualifier d'implicites lorsqu'il s'agit de connaissances intuitives utilisées par le système 1, et d'explicites lorsqu'il s'agit de connaissances rationnelles utilisées par le système 2. Cette hypothèse est inspirée de la théorie du « dual-process » comme formulée par Seymour Epstein [Epstein 1973] et Daniel Kahneman [Kahneman 2011].

Hypothèse 2 : La deuxième hypothèse que nous intégrons porte sur la hiérarchie des connaissances. Selon cette hypothèse il y aurait une certaine prévalence des connaissances implicites sur les connaissances explicites. Les connaissances implicites servent à la fois de briques fondamentales à la définition des connaissances explicites, et également d'amorces à la constitution de raisonnements rationnels. En d'autres termes, sans connaissances implicites il ne peut y avoir de connaissances ou de raisonnement explicites. Cette hypothèse est inspirée de la théorie des connaissances de Baruch Spinoza [Spinoza 1677].

Hypothèse 3 : La troisième hypothèse relève de l'existence d'un système de mesure et d'évaluation des états du corps. Ce système guide les processus cognitifs et sert de boussole aux comportements de l'individu. Cette hypothèse est inspirée du conatus ou théorie des Affects de Baruch Spinoza [Spinoza 1677] et des Marqueurs Somatiques de Antonio Damasio [Damasio 2003].

Hypothèse 4 : La quatrième hypothèse porte sur les formes fondamentales des connaissances. Selon cette hypothèse les notions d'espace et de temps sont les formes fondamentales des connaissances intuitives. Les connaissances rationnelles quant à elles ont la notion de concept comme forme fondamentale. Hypothèse inspirée par l'œuvre « Le monde comme volonté et comme représentation » de Arthur Schopenhauer [Schopenhauer 1888].

Hypothèse 5 : La cinquième hypothèse consiste à dire que les connaissances utilisées par la cognition, qu'elles soient implicites ou explicites, sont construites par l'individu au cours de son expérience corporelle avec l'environnement. Hypothèse inspirée de la notion d'émergence que l'on retrouve dans la partie « Le monde comme représentations » du livre de [Schopenhauer 1888] et la description de la notion d'énaction de Francesco Varela [Varela et al. 1993].

Hypothèse 6 : La sixième hypothèse porte sur la nature du langage. Selon cette hypothèse l'acte de langage est une succession de comportements moteurs retranscrivant des signes. En tant que comportement complexe du corps, l'acte de langage est une production automatique inconsciente. Hypothèse inspirée des travaux sur le langage et les symboles de Sigmund Freud [Freud 1905] et Carl Jung [Jung 1964].

Hypothèse 7 : La septième hypothèse porte sur l'interaction entre deux dimensions d'activité cognitive. Un premier niveau local, et un second niveau global. La multitude

d'activités locales donnant lieu à la propagation d'état globaux, diffusés et accessibles à l'ensemble de la cognition et dont le contenu pourrait être considéré comme le contenu de la conscience. Hypothèse inspirée de la théorie du « Global Workspace theory » formulée par Baars [Baars 1993] et par Dehaene et Changeux [Dehaene & Changeux 2005].

Hypothèse 8 : La huitième hypothèse relève de l'existence d'un principe dynamique faisant référence à la plasticité cérébrale. Selon ce principe deux modules cognitifs simultanément actifs tendent à créer une connexion. Hypothèse inspirée de la théorie de Hebb [Hebb 2005] et de l'apprentissage Hebbien, méthode d'intelligence artificielle qui en est inspirée [Kempter et al. 1999].

Hypothèse 9 : Enfin la dernière hypothèse que nous sélectionnons porte sur l'existence d'un principe dynamique dit « de renforcement » qui guide l'apprentissage. Hypothèse inspirée des théories du conditionnement de Pavlov [Pavlov 2010] et de l'ensemble de la branche de l'intelligence artificielle qui étudie l'apprentissage par renforcement (Reinforcement learning) [Sutton & Barto 2018].

3.2 IPSEL : le modèle fonctionnel

Nous formalisons un système de traitement de l'information comme une entité possédant des ressources, capable de recevoir des entrées et de produire des sorties. L'information qui entre dans le système est nommée *entrée*. Cette information *entrée* est ensuite traitée par le système via l'utilisation d'un second volume d'informations que nous nommons *ressources.* De ces traitements, un troisième volume d'information est produit, nous le nommons *sorties*.

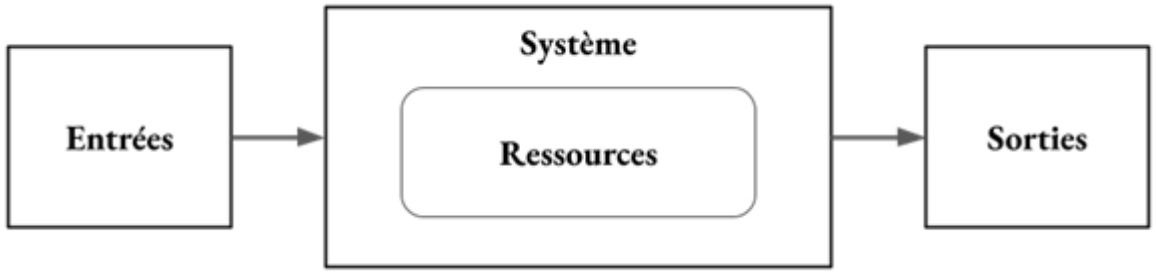

***Figure 3.2-1** : Forme générique d'un système de traitement de l'information.*

Un système de traitement de l'information (STI) est un mécanisme qui transforme un volume d'information en *entrée* en un volume d'information en *sortie,* par la mise en œuvre de traitements déterminés par un volume d'information qui sert de *ressources*.

En suivant ce formalisme, nous définissons également un système de traitement de l'information intelligent comme un STI dont les *ressources* peuvent être modifiées de telle sorte que le système améliore ses performances relativement à l'accomplissement d'un *objectif*.

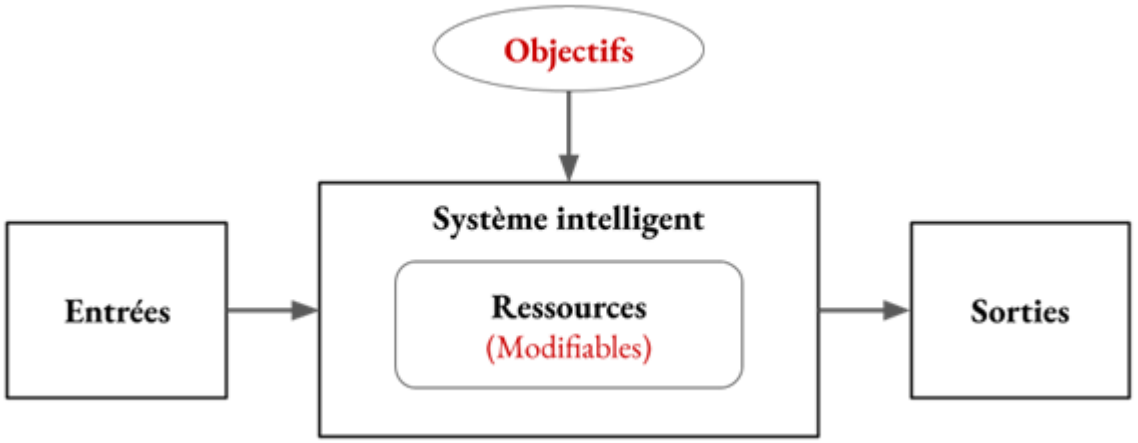

***Figure 3.2-2** : Forme générique d'un système de traitement de l'information intelligent.*

Le sujet que nous cherchons à modéliser en utilisant ce formalisme est la cognition humaine. En reprenant les hypothèses que nous avons choisies d'intégrer dans notre modèle, il s'agit de représenter l'architecture cognitive d'un agent autonome capable d'apprendre et de construire des connaissances implicites et explicites. Ces connaissances lui permettent de mettre en œuvre des comportements réactifs et automatiques mais aussi des comportements délibérés et complexes pour accomplir ses objectifs en s'adaptant à son environnement.

Si il fallait mobiliser la nomenclature employée dans le domaine de l'intelligence artificielle, nous pourrions dire que le système dont nous présentons l'architecture est :

1. Intelligent [*system*] (Capable d'apprendre)
2. Self-Supervised [*learning*] (Qui établit lui-même ses objectifs)
3. Emergent [*dynamism*] (Dont les structures internes s'auto-organisent)
4. Hybrid [*architecture*] (Qui combine les approches connexionniste et symbolique)

5. Statistical learning [*ability*] (Capable de découvrir et d'utiliser des corrélations statistiques)
6. Causal reasoning [*ability*] (Capable de découvrir et d'utiliser des logiques de causalités)

Ce qui correspondrait, une fois l'ensemble des propriétés concaténées, a :

➔ Self-Supervised Intelligent system with an Emergent Hybrid architecture capable of Statistical learning and Causal reasoning. (SSIEHSlCr)

Cependant nous nommerons notre modèle par un acronyme plus court et plus facile à employer :

➔ **Information Processing System with Emerging Logics (IPSEL)**

Ce choix a pour volonté de mettre en avant l'élément différenciant de notre proposition à savoir : la possibilité, pour un système de traitement de l'information, de faire émerger une capacité de raisonnement logique.

Pour mettre en œuvre cette capacité, le modèle IPSEL possède des caractéristiques spécifiques comme la décomposition de son architecture en 3 niveaux de traitement de l'information, l'intriquement de 2 de ces niveaux, et l'existence d'un module d'évaluation intrinsèque.

3.2.1 Architecture cognitive IPSEL

Le modèle IPSEL est une architecture cognitive. Cette architecture décrit l'articulation de 6 modules fonctionnels (OS, OE, S0, S1, S2, E) qui s'échangent des informations pour réaliser la cognition d'un agent.

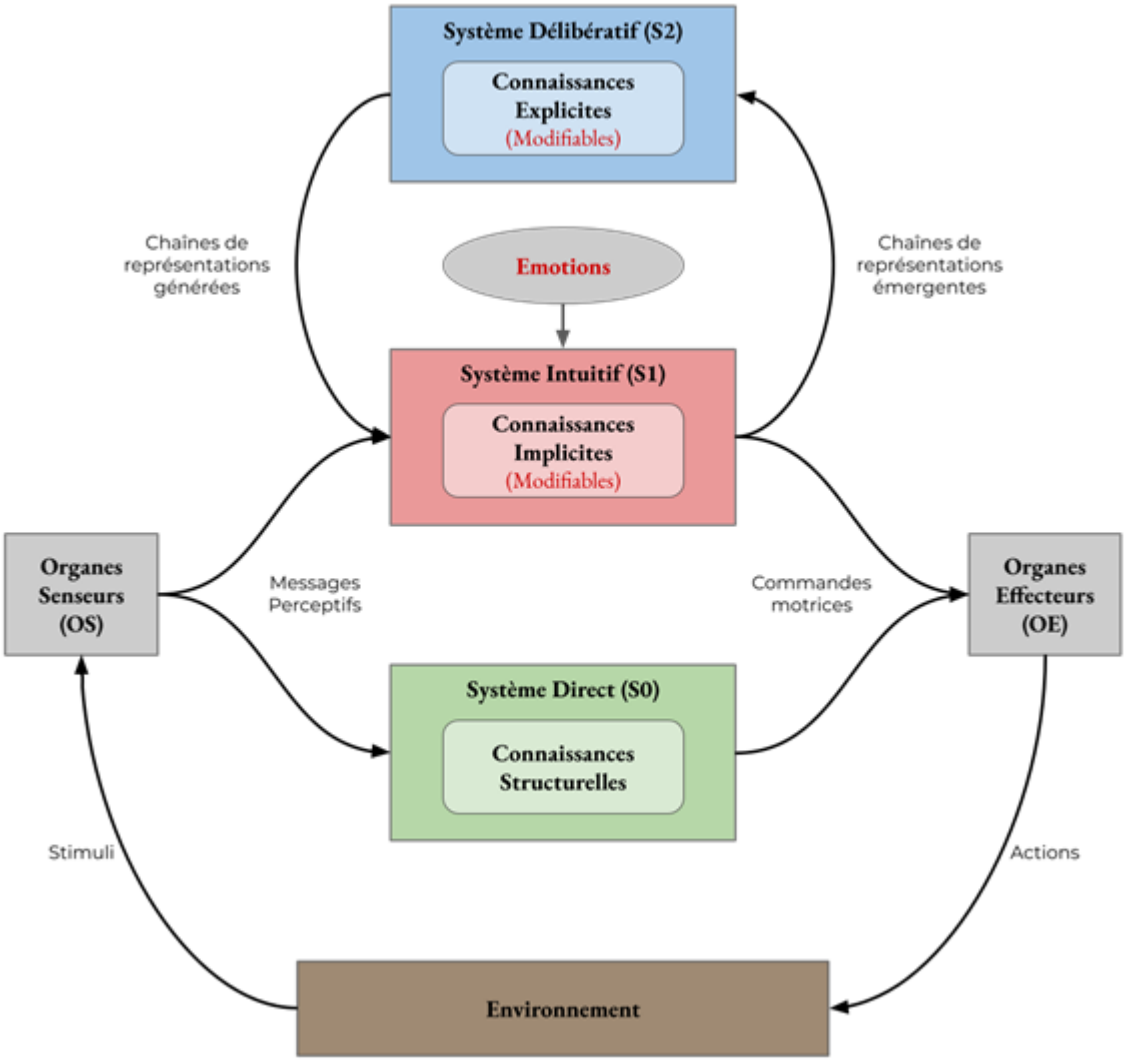

***Figure 3.2-3** : Architecture Cognitive I.P.S.E.L (Information Processing System with Emerging Logics).*

Les *organes sensori-moteur* (OS et OE) : Les *organes senseurs* (OS) reçoivent de l'information provenant de l'environnement. Nous nommons ces informations *stimuli*. Les *organes senseurs* transforment les *stimuli* en informations interprétables nommées *messages perceptifs*. A l'autre extrémité de la boucle agent-environnement, les *organes effecteurs* (OE) reçoivent des informations nommées *commandes motrices*, et les transforment en *actions* à opérer dans l'environnement. Les *organes senseurs* et les *organes effecteurs*, que nous regroupons sous le terme *organes sensori-moteurs*, sont les interfaces qui permettent à l'agent de percevoir son environnement et d'y agir, en mettant en œuvre des actions qui, prisent en tant qu'ensembles, forment les comportements de l'agent.

Sur la figure *3.2-3*, les organes OS et OE sont symbolisés chacun par un unique rectangle. Néanmoins, ils représentent potentiellement une multitude d'organes différentiables les uns des autres. Dans le cas où l'agent possède plusieurs organes réagissant, ou produisant différentes formes de *stimuli*, respectivement d'*actions*, on dira que OS, respectivement OE, sont multimodaux.

Le *Système Direct* (S0) : Il reçoit des *messages perceptifs* provenant des *organes senseurs* (OS) et utilise des ressources que nous nommons *connaissances structurelles* pour effectuer ses traitements. Il produit alors des *commandes motrices* envoyées aux *organes effecteurs* (OE) auxquels il est relié. Les *connaissances structurelles* sont des ressources fixées lors de la conception de l'agent. Le *Système Direct* (S0) ne réalise pas d'apprentissage, cela signifie que ses ressources ne peuvent pas être modifiées durant la vie de l'agent, ou alors seulement sur un principe d'érosion ou de dégradation non intentionnelle. Les *actions* commandées par le *Système Direct* composent des comportements que nous nommerons *végétatifs*.

Le *Système Intuitif* (S1) : Il reçoit trois types d'informations en *entrée*, (1) des *messages perceptifs* provenant des *organes senseurs* (2) des *chaînes de représentations générées* provenant du *Système Délibératifs* et (3) un signal objectif provenant du corps de l'agent que nous avons nommé *émotions*. Pour traiter ces informations et produire ses *sorties*, le *Système Intuitif* utilise des ressources modifiables nommées *connaissances implicites*. En utilisant ces différents volumes d'information, le *Système Intuitif* produit un premier type de *sortie* ; des *commandes motrices* à destination des *organes effecteurs*, qui mettront en œuvre des ensembles d'actions constituant les *comportements intuitifs*. L'activité du Système Intuitif produit également un deuxième type de *sortie* : les chaînes de représentations émergentes à destination du *Système Délibératif*.

Le *Système Délibératif* (S2) : Il reçoit des informations provenant du S1, elles sont structurées sous la forme de *chaînes de représentations émergentes*. Pour effectuer ses traitements le *Système Délibératif* utilise des ressources modifiables que nous appelons *connaissances explicites*. Le traitement des *chaînes de représentations émergentes* via l'utilisation des *connaissances explicites* permet au S2 de produire de nouvelles chaînes de représentations que nous appelons alors *chaînes de représentations générées*, à destination du *Système Intuitif*. Le *Système Délibératif* ne produit donc pas directement de *commandes motrices*. Nous dirons que les comportements engendrés par les traitements du S2 sont des

comportements de délibération, mais ils ne sont pas directement observables depuis l'environnement car ils ne commandent pas directement les *organes effecteurs*.

Les *émotions* (E) : L'*émotion* est un signal qui est produit par un module d'évaluation interne. Considéré de manière discrète l'*émotion* est un volume d'information qui représente une évaluation de l'état de l'agent par rapport à des états objectifs. Ce volume d'informations sert de fonction objective auto-générée guidant directement l'apprentissage, c'est-à-dire la modification des ressources du *Système Intuitif* et indirectement l'apprentissage du *Système Délibératif*.

3.2.2 Description fonctionnelle

Du point de vue fonctionnel, chaque système de traitement remplit un rôle spécifique, eu égard à ses caractéristiques particulières que nous allons maintenant détailler.

Le ***Système Direct***, et les *comportements végétatifs* qu'il produit, assurent la réalisation d'un certain nombre d'*actions* relatives à la sécurité et à l'intégrité de l'agent. Les *comportements végétatifs* sont mis en œuvre rapidement, et indépendamment des traitements des autres systèmes, grâce aux liens directs que le S0 trace entre certains *organes senseurs* et certains *organes effecteurs*. De plus, les ressources utilisées, *les connaissances structurelles*, sont initialement présentes dans l'agent en tant que ressources innées et ne subissent pas de modification au cours de la vie de ce dernier. Ces caractéristiques assurent la rapidité, l'indépendance et la stabilité des traitements du *Système Direct*. Les *comportements végétatifs* héritent de ces propriétés et sont donc adaptés pour assurer les *actions* qui doivent être mises en œuvre rapidement et dans tous les cas de figure tels que celles liés à la sécurité et à l'intégrité du corps de l'agent. La contrepartie est que ces *comportements végétatifs* ne permettent pas à l'agent de s'adapter à son environnement : puisque les ressources du S0 ne sont pas modifiables, aucune modulation des traitements qu'elles déterminent n'est possible.

L'adaptation à l'environnement est l'apanage du ***Système Intuitif***. Par sa capacité à construire et à modifier ses ressources, le S1 peut moduler les *comportements intuitifs* qu'il produit. Durant la vie de l'agent, les ressources du *Système Intuitif* se modifient de manière à représenter l'historique des associations [*messages perceptifs - commandes motrices - émotions].* Ces associations constituent alors les *connaissances implicites* et sont des ressources acquises. Elles seront mobilisées ultérieurement pour produire automatiquement

les *commandes motrices* les plus pertinentes par rapport aux *messages perceptifs* reçus. C'est-à-dire les *actions* qui, par le passé, ont engendré des *émotions* positives compte tenu des *messages perceptifs* considérés. Cet apprentissage est une généralisation des expériences passées, il suit un mécanisme inductif et s'apparente à de l'apprentissage statistique.

L'adaptation que réalise le *Système Intuitif* n'est cependant pas parfaite. Le mode d'acquisition des *connaissances implicites*, par généralisation des expériences passées, laisse l'agent sans connaissances face aux situations pour lesquelles il n'a pas d'expérience. Cette dépendance aux expériences passées rend également le *Système Intuitif* sensible à la représentativité de son expérience qui peut éventuellement biaiser les connaissances qui en sont induites. Qui plus est, l'émission des *comportements intuitifs* sur un seul mode réactif, perception/action, ne permet pas la planification de bénéfices émotionnels différés. A chaque instant, les *commandes motrices* sélectionnées automatiquement sont celles qui maximisent l'état escompté de l'agent dans le présent, or il y a des situations où une succession d'actions sous-optimales permettrait à l'agent d'atteindre de meilleurs états finaux dans le futur. Pour synthétiser ce point nous dirons que le *Système Intuitif* est fondamentalement ancré dans le présent et ne permet pas à l'agent de se projeter temporellement.

Le ***Système Délibératif*** a pour fonction de mitiger les défauts du *Système Intuitif* en permettant cette capacité de projection. Pour cela, le *Système délibératif* doit disposer d'un modèle causal représentant l'environnement, l'agent et leurs dynamiques. Pour parler de l'ensemble {agent + environnement} nous utiliserons désormais le terme *monde*. Le modèle causal est une représentation du *monde* en tant qu'ensemble d'objets ayant des caractéristiques perceptives et des propriétés dynamiques concernant l'évolution de ces caractéristiques. Ce modèle du *monde* (en anglais « *world-model* ») est représenté par les *connaissances explicites* et constitue la ressource du *Système Délibératif*. Ces connaissances sont construites au cours de la vie de l'agent à partir des *connaissances implicites*.

Une fois que les ressources du *Système Délibératif* se sont construites, au cours du fonctionnement de l'agent, le S2 reçoit des *chaînes de représentations émergentes* provenant de l'activité du *Système Intuitif*. Ces *chaînes de représentations émergentes* représentent une partie de l'état du *monde*, une partie sur laquelle est portée une certaine attention et que le S1 envoie au S2 pour que ce dernier les traite de manière plus approfondie. Grâce à ses connaissances explicites représentant la dynamique du *monde*, le *Système Délibératif* traite

les *chaînes de représentations émergentes* et déduit de nouvelles *chaînes de représentations* qui représentent des états simulés du *monde*, passés ou futurs. Ces *chaînes de représentations* ainsi *générées* par le *Système Délibératif* servent de nouvelles sources d'information à destination du *Système Intuitif*. Elles ont le potentiel d'informer, d'inhiber, de suggérer ou de planifier les comportements du *Système Intuitif* qui y réagit automatiquement. Le *Système Délibératif* ne commande pas directement les actions de l'agent, il agit comme un système d'aide à la décision pour le *Système Intuitif*.

Entrées	Système [Ressource]	Sorties	Comportements	Fonction
Stimuli	Organes senseurs (OS)	Messages perceptifs		Percevoir l'environnement
Commandes motrices	Organes effecteurs (OE)	Actions		Agir dans l'environnement
Messages perceptifs	Système Direct (S0) [Connaissances structurelles]	Commandes motrices	Végétatifs	Assurer la sécurité et l'intégrité du corps
Messages perceptifs, Chaînes de représentations générées, Émotions.	Système Intuitif (S1) [Connaissances implicites]	Commandes motrices, Chaînes de représentations émergentes	Intuitifs	Contrôler et adapter automatiquement les comportements du corps avec l'environnement, initier les traitements du S2
Chaînes de représentations émergentes	Système Délibératif (S2) [Connaissances explicites]	Chaînes de représentation générées	Délibératifs	Aide à la décision pour le S1 via la simulation des états du corps et de l'environnement.
	Corps de l'agent	Emotions (E)		Evaluation de l'état du corps

***Figure 3.2-4** : Tableau récapitulatif des modules de l'architecture IPSEL et de leurs fonctions.*

3.2.3 Intrication des systèmes intuitif et délibératif, dialogue intérieur.

Au cours de la vie de l'agent, les ressources du *Système Intuitif* sont modifiées de telle sorte qu'elles représentent les associations passées entre l'état de l'environnement réel, l'état de l'agent, l'état du *monde* simulé que nous nommerons alors *pensée* de l'agent, et les commandes motrices. Ces ressources appelées *connaissances implicites* sont construites sur un mode inductif, par généralisation des associations passées. Ce mécanisme d'apprentissage s'apparente à un apprentissage statistique.

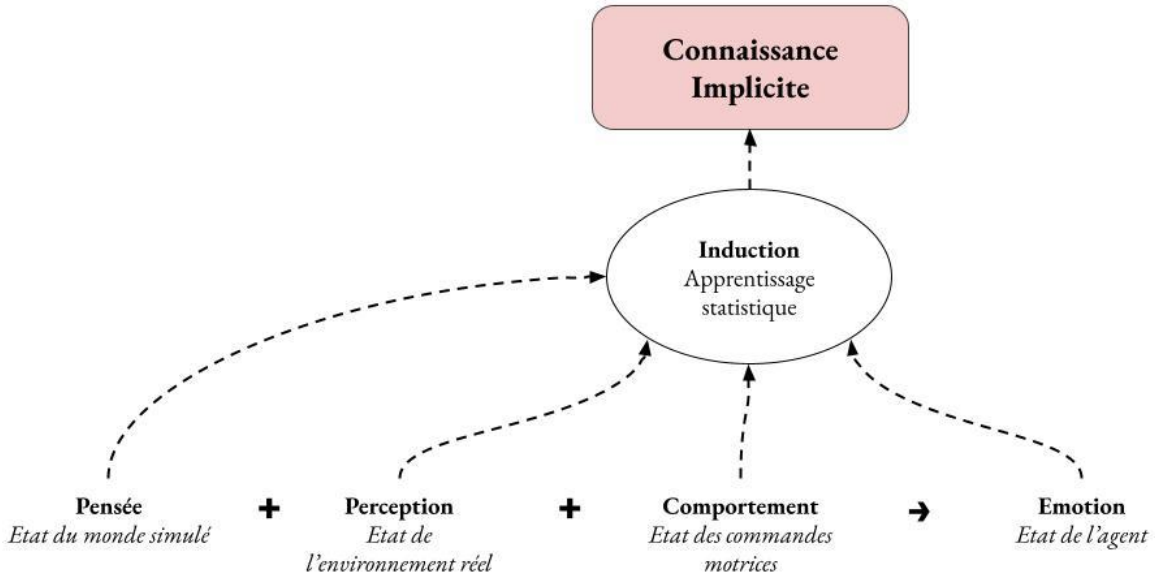

Figure 3.2-5 *: Construction des connaissances implicites.*

Toujours au cours de la vie de l'agent, les *connaissances implicites* s'auto-organisent pour former des structures ayant des relations de composition et de causalité les unes avec les autres. Ces structures émergentes forment les *connaissances explicites*. Toutes les *connaissances explicites* peuvent être exprimées sous la forme de composition de *connaissances implicites*. En revanche, toutes les *connaissances implicites* ne peuvent pas être exprimées sous la forme de *connaissances explicites*. Seules les *connaissances implicites* qui appartiennent et constituent des structures émergentes peuvent être exprimées sous la forme de *connaissances explicites*. Les *connaissances explicites* sont en quelque sorte une retranscription de certaines connaissances implicites exprimées dans un format explicite et relationnel. On pourrait également voir les *connaissances explicites* comme un format compressé (au sens de compression de l'information) des *connaissances implicites*.

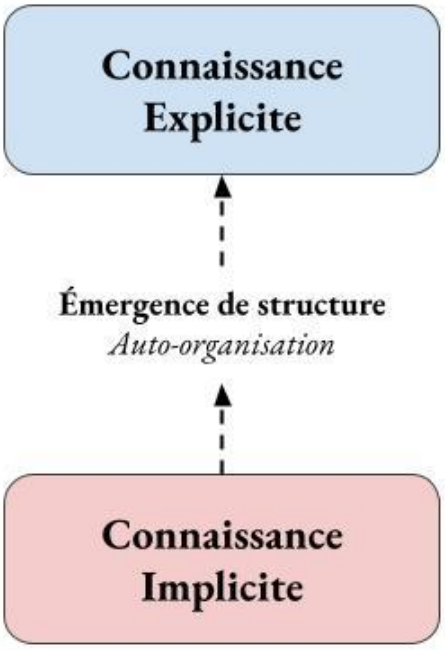

***Figure 3.2-6** : Construction des connaissances explicites.*

Les *connaissances explicites* représentent les objets du *monde* sous une forme structurée. Les relations de compositions et de causalité entre les *connaissances explicites* représentent la dynamique des objets du *monde*. Grâce à ces relations, le *Système Délibératif* peut déduire de nouvelles *représentations explicites* via un mécanisme d'inférence logique. Une *représentation explicite* peut être retranscrite, ou convertie, en une composition de *connaissances implicites,* et cela est permis car les c*onnaissances explicites* sont des structures de *connaissances implicites*. Comme nous l'avons décrit précédemment, les *connaissances implicites* intègrent les émotions en tant qu'état associé de l'agent. Par ce biais, les *chaînes de représentations générées* peuvent être indirectement évaluées par rapport aux *émotions* associées à leur transposition en *connaissances implicites*.

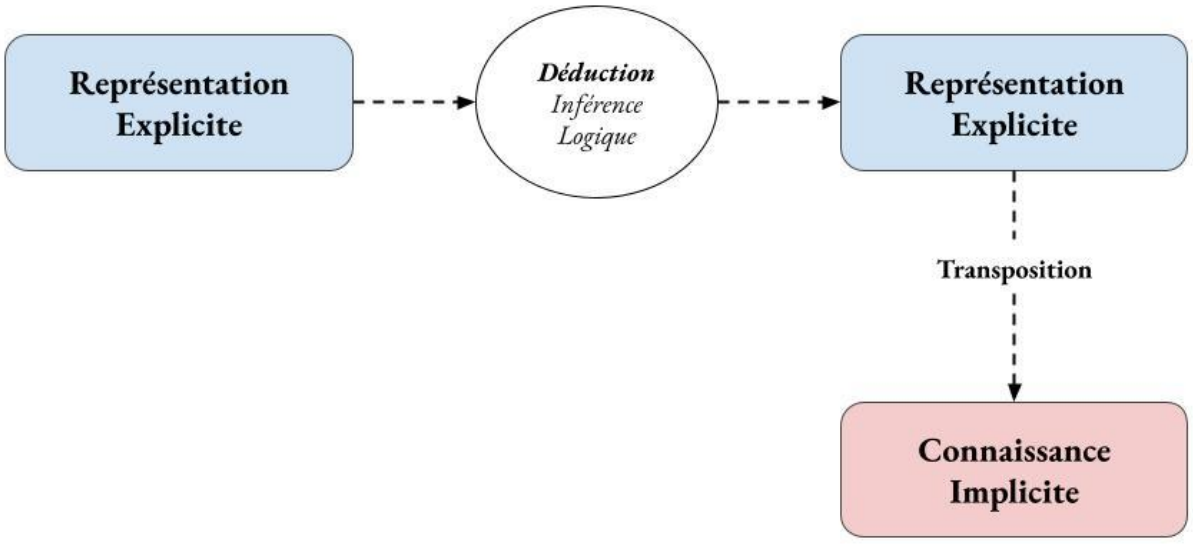

***Figure 3.2-7** : Déduction de représentation explicite et transposition en connaissances implicites.*

Ce rapport qui existe entre les *connaissances implicites* et *explicites* permet aux *systèmes Intuitif* et *Délibératif* d'échanger de l'information. En outre, les ressources du *Système Délibératif* sont des constructions qui nécessitent une phase d'auto-structuration des ressources du *Système Intuitif*, on dit alors que le *Système Délibératif* émerge de l'activité du *Système Intuitif*.

Nous caractérisons ces systèmes comme intriqués, au sens où le *Système Délibératif* ne peut pas exister sans le *Système Intuitif* puisque ses ressources sont des structures de *connaissances implicites*, et que l'initiation de ses traitements nécessite l'initiation d'une première *chaîne de représentations* fournie automatiquement par le S1. Qui plus est, les productions du *Système Délibératif* n'ont d'utilité que si le S1 les prend en considération pour ses propres traitements. Les *Système Intuitif* et *Système Délibératif* fonctionnent en coopération continue et sont interdépendants. Symboliquement, on dira que les comportements de l'agent sont le résultat d'un dialogue intérieur entre S1 et S2.

3.2.4 Description synthétique et éléments différenciants.

Nous donnons maintenant une description synthétique d'un agent dont la cognition suit l'architecture IPSEL, en insistant plus particulièrement sur les éléments qui différencient le plus notre modèle des autres propositions de la sorte.

Un agent IPSEL a vocation à optimiser une fonction interne qui évalue les états de son corps. Pour ce faire, il dispose de connaissances innées et de mécanismes pour en acquérir de

nouvelles (*nature and nurture*). Les premières connaissances que l'agent acquiert ont des structures associatives simples, elles sont induites à partir de l'historique de ses interactions avec l'environnement et lui permettent de déployer des comportements d'adaptations instinctifs. Au fil du temps, les structures de sa mémoire s'organisent pour former des connaissances plus complexes. L'utilisation de ces connaissances structurées permet à l'agent de penser le *monde* et de moduler ses comportements instinctifs.

Les spécificités de l'agent IPSEL sont qu'il construit son propre modèle du *monde* à partir de son expérience. Ce faisant, il lui est permis d'utiliser ce modèle du *monde* pour raisonner, mais ces raisonnements ne sont que informatifs et ne contrôlent pas directement les actions de l'agent. Dans la plupart des autres architectures, le module responsable des raisonnements logiques contrôle directement les actions de l'agent.

Avec un autre vocabulaire, on pourrait dire que le modèle IPSEL architecture un système connexionniste ayant vocation à construire et utiliser un modèle implicite du *monde*, avec un système symbolique qui utilise un modèle logique. Cependant le système logique nécessite une définition de ses éléments et des règles d'inférence, ces définitions sont construites par le système connexionniste. Cela différencie l'architecture IPSEL par rapport aux autres architectures hybrides où la logique du système symbolique est formatée par l'Homme, comme dans le cas des systèmes experts par exemple.

3.3 Hipsel : une machine mettant en œuvre l'architecture IPSEL

Afin de donner maintenant une intuition sur une potentielle mise en œuvre des principes de l'architecture IPSEL, nous décrivons le fonctionnement d'une machine hypothétique comme exemple d'instanciation du modèle. Nous nommerons cette machine Hipsel pour « Hypothetical-IPSEL ». IPSEL sans « H » est un modèle fonctionnel qui pourrait être implémenté de multiple manière. Hipsel avec un « H » est une machine hypothétique, c'est une proposition de ce que pourrait être l'une des manières d'implémenter le modèle IPSEL.

Hipsel possède un corps dont certaines parties accomplissent des fonctions bien particulières. Les *organes senseurs* (illustrés par des « radars » sur la figure *3.3-1*) sont des parties du corps qui, en interaction avec l'environnement ou d'autres parties du corps, produisent des *messages perceptifs* (illustrés par des sphères jaunes sur la figure *3.3-1*). Les *messages perceptifs* sont acheminés à d'autres parties du corps qui, prises dans leur ensemble, forment la *cognition* (illustrée par un cube bleu aux arêtes arrondies sur la figure *3.3-1*). La

cognition se sert des *messages perceptifs* et d'autres ressources pour produire des messages appelés *commandes motrices*. Les *commandes motrices* sont ensuite acheminées aux *organes effecteurs* (illustrés par des engrenages sur la figure *3.3-1*). Activés par les commandes, les *organes effecteurs* mettent le corps d'Hipsel en mouvement via le contrôle moteur des parties du corps auxquels ils sont reliés.

Le corps dans son ensemble émet un signal qui rend compte de son état, ce signal est appelé ici *État émotionnel* (illustré par la fenêtre multi-couleur sur la figure *3.3-1*), il est visible de la cognition et constitue l'une de ses ressources. Le rôle de la cognition est de coordonner les mouvements du corps avec l'environnement tout en régulant les *états émotionnels* du corps.

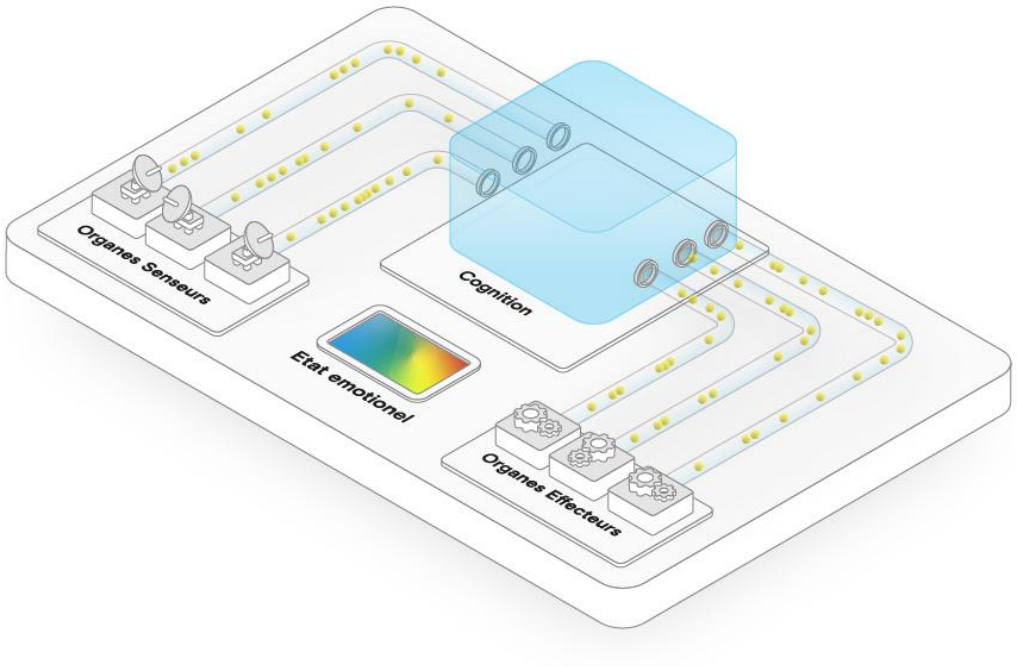

***Figure 3.3-1** : Les 4 parties de la machine et leurs communications. (i) les organes senseurs, (ii) Les systèmes de traitement regroupés en tant que Cognition, (iii) les organes effecteurs. (iv) le module d'évaluation se manifestant ici à travers des « états émotionnels ».*

La cognition est réalisée par un ensemble de parties du corps activables et qui agissent comme une mémoire. L'information provenant des *organes senseurs* active certaines structures de la mémoire que l'on appelle alors *représentations* (illustrées par un cube gris aux arêtes droites sur la figure *3.3-2*), car leurs activités représentent l'activité des autres

parties du corps auxquelles ces structures sont reliées et qui leur ont propagé un signal d'activation.

Une fois activées, les *représentations* peuvent activer d'autres structures du corps auxquelles elles sont reliées. Ces autres structures sont soit d'autres *représentations*, c'est-à-dire d'autres structures de la mémoire, soit des *organes effecteurs*. Dans ce dernier cas, on dira que la cognition a produit une *commande motrice*.

Pour s'activer mutuellement, les *représentations* partagent des liens que nous appelons *logiques* (illustrés par un trait gris reliant deux cubes sur la figure *3.3-2*). Une *logique* entre une *représentation* R1 et une *représentation* R2 incarne une règle de probabilité qui donne la probabilité que R2 soit activée sachant que R1 est active. Ces liens sont orientés et représentent une relation de causalité entre les *représentations*. Si la probabilité que R1 soit active, sachant que R2 est active, est nulle, et inversement si la probabilité que R2 soit active, sachant que R1 est active, est également nulle, alors les *représentations* R1 et R2 ne partagent aucune *logique*, aucun lien, leurs activités sont indépendantes.

Une *connaissance* est un ensemble de *représentations* liées par des *logiques* (illustrée par un ensemble de cubes ayant des liens entre eux sur la figure *3.3-2*).

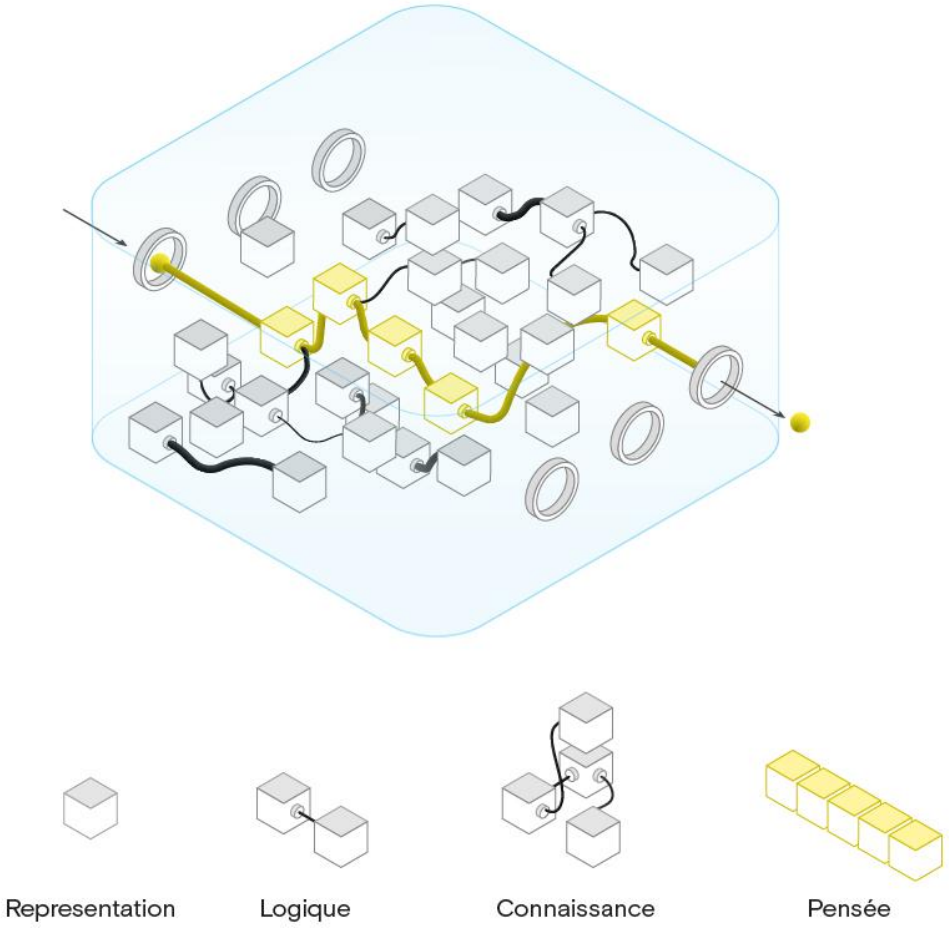

***Figure 3.3-2** : Cognition de Hipsel, illustration d'une représentation, d'un lien représentant une logique, d'une connaissance comme un ensemble de représentations partageant des logiques et d'une pensée comme phénomène d'activation simultané d'une séquence de représentations.*

Une *pensée* est un phénomène au cours duquel plusieurs *représentations* sont activées séquentiellement (illustrées par un ensemble de cubes jaunes ayant des liens entre eux sur la figure *3.3-2*). Ici le terme *pensé* ne fait pas nécessairement référence à une pensée conscientisée au sens anthropomorphique, il s'agit simplement d'une chaîne de *représentations* actives plus ou moins simultanément.

Les *états émotionnels* sont des états du corps évalués à partir de plusieurs composantes (illustrés par un cylindre sur la figure 3.3-3) qui agissent comme des mesures. Un *état émotionnel* est une composition particulière de la valeur de ces diverses mesures. Les états émotionnels possèdent une valence (illustrée par une étiquette sur la figure 3.3-3). La

valence informe sur la désirabilité de l'*état émotionnel* par rapport à des états objectifs, certaines gammes d'*états émotionnels* sont désirables et d'autres indésirables. Cette attribution de valence pourrait avoir émergée d'un mécanisme évolutionniste ou avoir été préétablie lors de la phase de conception de la machine.

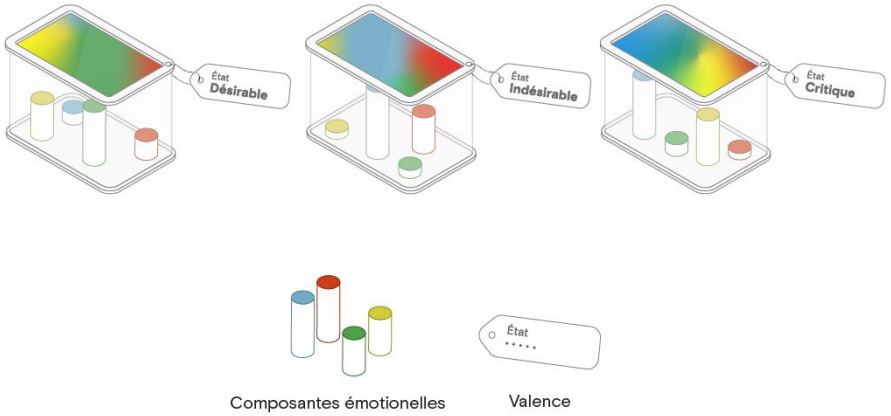

***Figure 3.3-3** : Composantes et valences des États émotionnels du corps d'Hipsel.*

Les structures de la cognition sont dites « plastiques ». Cela signifie que leurs formes peuvent changer au cours du temps, leur permettant d'accomplir leur rôle de mémoire. Ces différenciations ne sont pas aléatoires mais dirigées de telles sortes qu'elles servent l'accomplissement des fonctions qui incombent à la cognition. Cette auto-organisation des structures de la cognition constitue le mécanisme d'apprentissage d'Hipsel. La fonction objective de cet apprentissage est l'*état émotionnel*, sa contrainte est la dépense énergétique. En effet, pour activer des *représentations* ou des *organes effecteurs*, l'agent consomme de l'énergie. L'amélioration des performances de la cognition réside alors dans le fait d'organiser les structures de la mémoire de telle sorte qu'elles contrôlent le corps de manière à surveiller les *états émotionnels* (c'est-à-dire les maintenir dans des états désirables et éviter les états indésirables) tout en gérant l'énergie qu'elle dépense pour le faire. Pour réaliser cet acte, les structures de la mémoire suivent un principe dynamique associatif, assimilable à un apprentissage par renforcement. Les structures qui sont activées simultanément tendent à créer des liens, et ce faisant créent des logiques et des *connaissances*. Si ces structures possèdent déjà un lien, il est renforcé. L'intensité du renforcement est alors proportionnelle à l'intensité de l'état émotionnel au moment où ces structures sont actives. Au cours du temps, les liens qui ne sont pas activés se dégradent. Ainsi les structures de la mémoire sont

soumises à un dynamisme qui permet la création, la modification et la suppression de connaissances (illustration *Figure 3.3-4*).

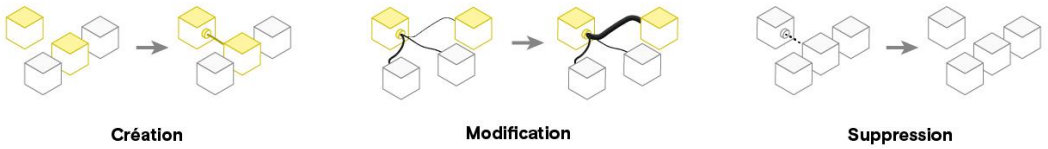

***Figure 3.3-4** : Dynamique des structures de la cognition d'Hipsel.*

En se connectant, les *représentations* forment de nouvelles structures de plus haut degré que nous pouvons considérer comme des représentations de *représentations*. Pour qualifier ces différents degrés de *représentation*, nous utilisons le terme *concret* pour parler d'une représentation de bas degré, qui est par exemple en lien direct avec un *organe senseur*. Et à l'inverse, une *représentation* de plus haut degré, qui représente d'autres représentations, sera qualifiée d'*abstraite* dans le sens où elle fait intervenir plus de structures, donc plus de *connaissances*. On peut également dire que si l'activation d'une *représentation* consomme de l'énergie et nécessite un certain temps pour que l'ensemble de la structure soit activé, alors les représentations *abstraites* consomment plus d'énergie et sont plus lentes à activer puisqu'elles impliquent plus de structures.

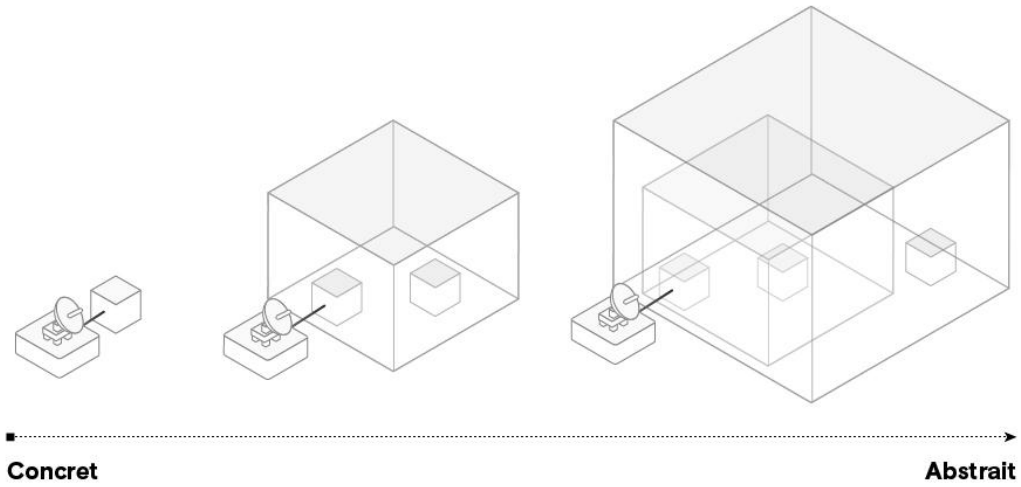

***Figure 3.3-5** : Illustration de l'encapsulation des représentations pour former des connaissances de plus haut degré d'abstraction (composition verticale).*

L'environnement dans lequel évolue le corps d'Hipsel n'est pas chaotique. Les phénomènes qui y ont lieu présentent un certain ordre dans le sens où ces évènements sont parfois récurrents, invariants ou symétriques. Nous dirons que l'environnement et sa dynamique suivent certains motifs, et que ces motifs sont donc retranscrits dans le flux des *messages perceptifs*.

L'auto-organisation des structures de la mémoire a pour but d'apprendre à représenter ces motifs en tant que *connaissances* utiles pour la coordination du corps avec l'environnement.

Dans le flux des *messages perceptifs*, les motifs symbolisent certains aspects invariants du flux. Ces invariances dans l'activité des *organes senseurs,* et donc des *messages perceptifs*, engendrent une répétition de l'activation des structures de la mémoire qui les représente. En conséquence des principes dynamiques, les structures de la mémoire qui représentent les éléments du motif finissent par se connecter et forment une nouvelle structure, une structure de structure ou *connaissance*, qui représente le motif dans son ensemble. Un motif est identifiable par rapport à un ou plusieurs critères qui servent de référentiel en fonction desquels s'exprime l'invariance qui le définit. Nous énonçons donc le principe général qui guide la construction de *représentation* comme suit :

- Étape 1 : Invariance de l'activité par rapport à un ou plusieurs critères → motif

- Étape 2 : Activation simultanée et récurrente, des structures représentant les constituants du motif → construction de liens entre les structures constituantes du motif.

N'étant plus indépendantes mais *logiquement* liées, les structures initialement isolées forment une nouvelle structure de plus haut degré, une *connaissance* plus abstraite qui représente le motif dans son ensemble.

A l'origine, un motif peut s'exprimer par rapport à trois critères fondamentaux. Une *représentation*, en tant que structure active de la mémoire, possède trois propriétés fondamentales; sa localisation dans le réseau de représentations (différenciation par rapport à l'espace), sa durée d'activation (différenciation par rapport au temps) et ses relations aux autres structures, notamment les structures qui causent son activation et les structures qui seront activées comme effets de son activation (différenciation par rapport à la causalité). Discriminer les représentations par rapport à leurs localités respectives permet d'initier une notion d'espace, nous dirons alors que s'expriment des motifs spatiaux. Discriminer les représentations par rapport à leurs durées d'activation respectives initie une notion de temps, nous dirons que s'expriment des motifs temporels. Discriminer les représentations par rapport aux causes et aux effets de leurs activations initie une notion de causalité, nous dirons que s'expriment des motifs causaux.

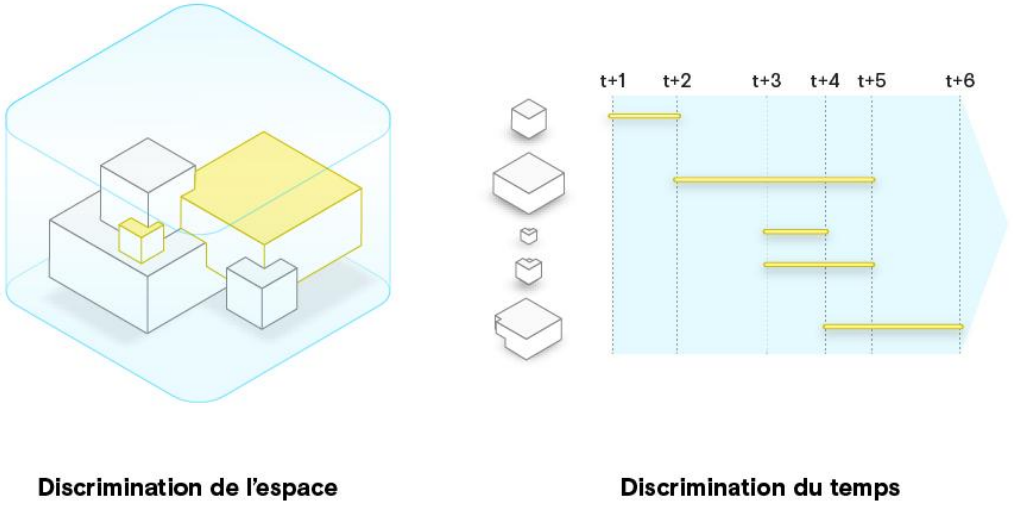

***Figure 3.3-6** : Illustration des formes fondamentales de l'espace et du temps des représentations. Critère de leurs discriminations les unes par rapport aux autres : leurs positions respectives dans l'espace et l'ordre de leurs activations dans le temps.*

L'association des critères d'espace, de temps et de causalité crée une nouvelle dimension dans laquelle des motifs peuvent s'exprimer, et donc être représentés. Nous les appellerons des *objets de perception*. Les *objets de perception* sont des motifs espace/temps/causalité invariants.

Discriminés par rapport aux critères fondamentaux de l'espace, du temps et de la causalité, les *objets de perception* ont des relations entre eux. Certaines de ces relations sont invariantes, d'autres sont proportionnelles, l'ensemble de ces relations entre les *objets de perception* forme une nouvelle dimension dans laquelle des motifs peuvent s'exprimer, et donc être représentés, toujours par application des principes dynamiques auxquels sont soumises les structures de la mémoire. Ces nouveaux motifs portant sur les relations entre les objets seront appelés *concepts*. Les concepts sont des motifs de relations inter-objets invariants.

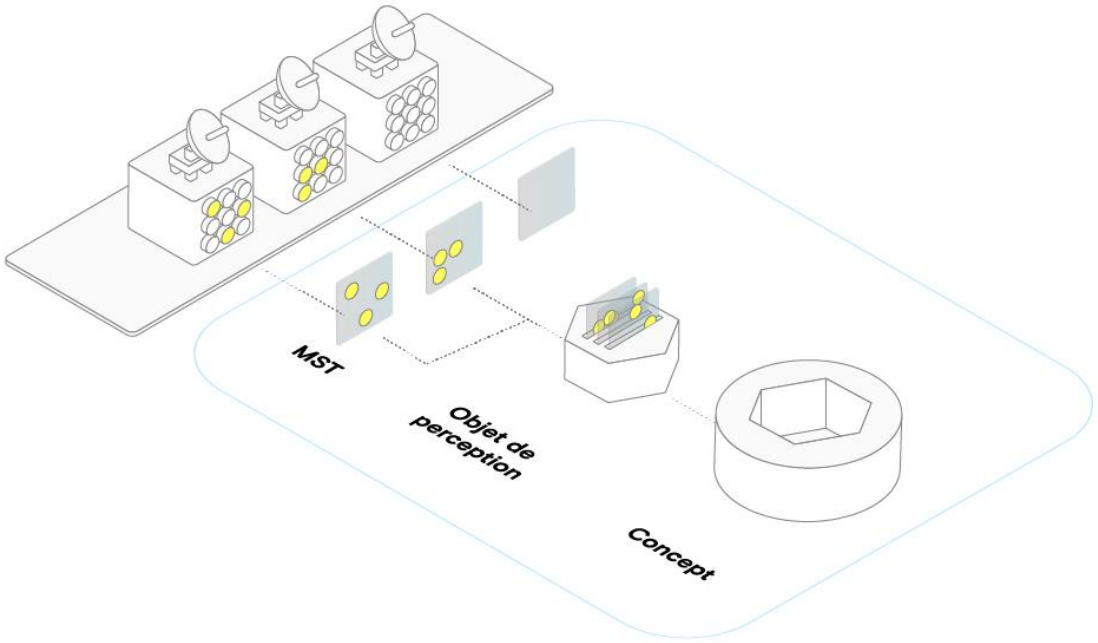

***Figure 3.3-7** : Illustration de l'association de représentation de même degré d'abstraction pour former des « motifs de motifs » (composition horizontale).*

Une *pensée* est un phénomène durant lequel un ensemble de *représentations* s'active en une séquence. Une *pensée* possède donc une durée. De manière générale, la cognition est le siège permanent d'une multitude de *pensées* de diverses formes, c'est-à-dire impliquant des *représentations* de divers degrés. Ce sont les *pensées* qui, en activant successivement les *représentations*, permettent aux activations engendrées par les *organes senseurs* de se transformer en *commandes motrices* à destination des *organes effecteurs*. Pour lier entrées et sorties, les *pensées* prennent un chemin particulier à travers différentes structures de *représentation* et qui est déterminé par les *logiques* qui relient ces structures. Les phénomènes de *pensée* sont donc le mécanisme de traitement de l'information, qui transforme les entrées en sorties en utilisant les ressources de la cognition. Ces ressources sont les connaissances innées, détermination originelle du réseau des structures de la mémoire, les connaissances acquises, les structures du réseau qui se sont construites en conséquence des principes dynamiques, et les *états émotionnels*, qui agissent comme une «signalétique» influençant la sélection de trajectoires dans le réseau.

Pour retrouver les trois systèmes de traitement de l'information de l'architecture IPSEL il faut identifier trois types de *pensées* :

Les pensées qui activent des *représentations* de premier degré, concrètes, et qui parcourent des chemins directs, pour lesquelles aucune autre trajectoire n'est possible. Indépendants et isolés des autres réseaux, ces chemins directs appartiennent aux structures innées et ne se ramifient pas avec d'autres réseaux. Ce type de *pensée* pourrait s'appeler *pensées corporelles*, elles correspondent au traitement du *Système Direct* de l'architecture ISPEL (illustrées par le chemin direct, n'impliquant aucun réseau, dans la figure 3.3-8). Puisque ces *pensées* font intervenir des chemins uniques et isolés, traversant des structures de bas degré, leur propagation est très rapide, consomme peu d'énergie et n'est pas modulable.

Les *pensées* qui font intervenir des structures plus complexes, des réseaux dans lesquels plusieurs chemins sont possibles, seront nommées *pensées intuitives* (illustrées par le réseau d'hexagones en bas au fond dans la figure 3.3-8). Dans ces réseaux, la sélection d'une trajectoire est modulée dans l'instant par l'influence des états émotionnels, et au cours de la vie de l'agent par le dynamisme des structures qui se modifient en fonction des interactions environnementales. Ce type de *pensée* correspond au traitement du *Système Intuitif* dans l'architecture IPSEL. On dira que ces *pensées* opèrent à des niveaux locaux.

Enfin les *pensées* qui font intervenir des *représentations* de haut degré, abstraites, seront nommées *pensées conceptuelles* et correspondent au traitement du *Système Délibératif* de l'architecture IPSEL (illustrées par le réseau d'hexagones situés en haut de la figure 3.3-8). Les *pensées conceptuelles* impliquent des structures de haut degré, des structures de structures, on dira que ces *pensées* opèrent à des niveaux globaux, et peuvent donc être mises en correspondance avec des pensées aux niveaux locaux, des *pensées intuitives*.

Nous définissons deux mécanismes qui effectuent ces correspondances entre les niveaux locaux et globaux : (i) La *conceptualisation* (C(r) sur la figure 3.3-8) où des connaissances locales sont mises en correspondance avec des connaissances globales, et (ii) L'objectivation (O(r) sur la figure 3.3-8) où des connaissances globales sont mises en correspondance avec des connaissances locales. Cela est rendu possible car les *représentations* de haut niveau sont construites à partir de *représentations* de plus bas niveau.

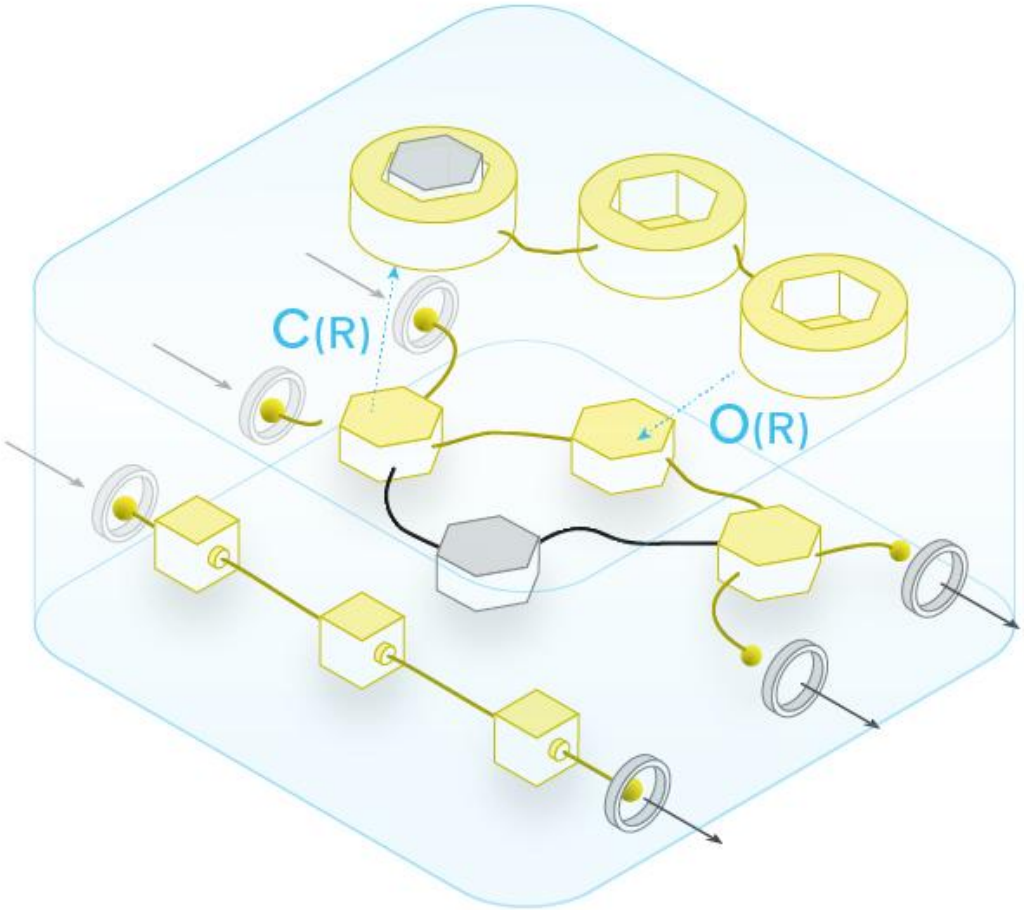

***Figure 3.3-8** : Illustration des différentes formes de pensées, relatives aux formes des représentations qu'elles mobilisent. Relations entre pensées intuitives et pensées délibérées par l'intermédiaire des mécanismes de conceptualisation et d'objectivation.*

La machine cognitive Hipsel ainsi décrite est donc un corps en interaction avec un environnement. Ces interactions activent certaines structures internes de la machine. Au cours de l'expérience de la machine, ces activations entraînent une auto-organisation des structures de la mémoire qui forment alors des connaissances acquises. Ces connaissances détermineront les *pensées*. Les *pensées* mettent le corps de la machine en mouvement. On distingue trois archétypes de *pensées* qui ont lieu dans la cognition de Hipsel. Les *pensées corporelles, intuitives et conceptuelles*. Les *pensées corporelles* utilisent des connaissances encodées dans les structures immuables du corps d'Hipsel. Elles ne sont pas sujettes à un apprentissage. Les *pensées intuitives* résultent de l'utilisation de connaissances qui ont été

construites au cours de l'expérience et représentent le corps et l'environnement en termes d'*objets de perception* associés à des *états émotionnels.* Elles permettent à Hipsel de moduler et de coordonner les comportements de son corps avec celui de l'environnement en utilisant son expérience de manière à rechercher et maintenir les états émotionnels considérés comme désirables. Les *pensées conceptuelles* utilisent des connaissances qui représentent le corps et l'environnement en termes de *concept*, c'est-à-dire des classes d'objets ayant des propriétés et des relations fonctionnelles avec d'autres *concepts*. Elles permettent à Hipsel de rationaliser son comportement en utilisant son savoir pour simuler des environnements passés ou futurs par déduction, et ainsi influencer la trajectoire des *pensées intuitives* qui contrôlent le corps.

3.4 Discussions et paradigme

Dans cette partie, pour continuer à transmettre l'intuition derrière le modèle IPSEL, nous conjecturons quelques propriétés qui pourraient être attribuées aux comportements humains et aux machines artificielles dans le cas où l'on accepte IPSEL comme modèle d'un appareil cognitif équivalent à celui de l'Homme. Nous espérons montrer que le modèle est à minima utile comme cadre théorique pour discuter ces éléments.

3.4.1 Lectures des comportements humains dans le cadre de la théorie IPSEL.

La théorie du « Dual-Process » propose de considérer la cognition humaine comme résultant de l'activité de deux systèmes, l'un basé sur l'expérience et le second sur la logique. Nous avons intégré cette théorie à l'architecture IPSEL comme présenté dans sa description fonctionnelle en partie 3.2 de ce chapitre. Lors de la présentation de la machine cognitive Hipsel, partie 3.3, nous avons également décrit comment le système d'aide à la décision basé sur la logique, le *Système Délibératif* (S2) pourrait émerger de l'auto-organisation des ressources du *Système Intuitif* (S1) basé sur l'expérience et les émotions. En raison de ce phénomène d'émergence ces deux systèmes sont intrinsèquement liés. Les *connaissances implicites* servent à construire les *connaissances explicites* et initient les *raisonnements explicites*, c'est l'influence du S1 sur S2. En retour les *raisonnements explicites* servent de nouvelles sources d'information pour les traitements du *Système Intuitif*, c'est l'influence du S2 sur le S1. Ce phénomène permet la perpétuation de cycles d'échanges d'information entre les systèmes 1 et 2. A chaque instant, les productions de la cognition sont le résultat d'une négociation entre ces deux systèmes qui s'influencent et communiquent lors d'un

« dialogue intérieur », d'une boucle ou d'un débat entre expérience et raisonnement, entre automatisme et délibération. Dès lors, il serait incomplet d'attribuer une production cognitive uniquement au système 1 ou au système 2 tant leurs traitements dépendent l'un de l'autre.

Nous en conjecturons un principe de dualité cognitive qui consiste à entrevoir toute production cognitive comme le fruit de multiples phénomènes qui, pour en simplifier l'analyse, peuvent être catégorisés en deux ensembles; les phénomènes cognitifs intuitifs et les phénomènes cognitifs délibérés. L'analyse d'un comportement physique ou d'un phénomène psychique devrait alors, pour prendre en compte le principe de dualité cognitive, considérer les facteurs intuitifs et les facteurs délibérés qui ont engendré l'élément à analyser.

Par exemple, l'analyse de la confiance de l'individu pourra se faire, par application du principe de dualité cognitive, via l'analyse de deux types de confiance : une confiance intuitive qui résulte de la mobilisation des *connaissances implicites*, de l'expérience et de l'état émotionnel de l'individu, et une confiance délibérée résultant de la mobilisation des *connaissances explicites* du sujet telles que son savoir logique et conceptuel concernant la situation et l'objet jugé. Cette décomposition permet de discriminer plusieurs cas de figure, comme le cas où les confiances intuitives et délibérées convergent et le cas où elles divergent. Cette décomposition permet également d'affiner les hypothèses concernant les actions à mener pour « réparer » un défaut ou un excès de confiance de l'individu, puisque du fait des différences de caractéristiques entre les systèmes 1 et 2, un défaut de confiance intuitive ne se mitige pas de la même manière qu'un défaut de confiance délibéré. Nous développerons la discussion de la confiance dans le chapitre suivant, partie 4.4, dans la discussion de l'IA de confiance.

De plus, dans le modèle IPSEL, les connaissances des systèmes 1 et 2 sont des ressources qui se sont organisées pour représenter implicitement les *messages perceptifs*, les *commandes motrices* et les *états émotionnels,* et pour représenter explicitement le corps, l'environnement et leurs dynamiques. Les formes de ces connaissances sont relatives aux modalités de perception de l'agent et leur contenu est déterminé par son expérience individuelle. Ainsi deux agents peuvent construire des connaissances similaires, mais elles ne peuvent jamais être considérées comme totalement équivalentes puisque leurs expériences et l'apprentissage qui en a résulté sont uniques et individuels.

Nous conjecturons un second principe, dit de subjectivité générale. Il est une conséquence de l'hypothèse d'émergence et du fait que les ressources utilisées par les systèmes *Intuitifs* et *Délibérés* soient des constructions. Il postule que toute représentation cognitive de l'agent est subjective en cela qu'elle a été construite par l'agent en fonction de son expérience individuelle. Lors de l'analyse d'un phénomène psychique ou d'un comportement physique, appliquer le principe de subjectivité générale consistera alors à préciser le sujet et les circonstances associées à l'élément analysé, et à ne jamais estimer que cet élément est une conséquence nécessaire, absolue et universelle. Ce principe est analogue à la position philosophique du relativisme.

Par exemple, pour l'attribution de la propriété « de confiance » aux systèmes utilisant l'intelligence artificielle, appliquer le principe de subjectivité générale consiste à affirmer qu'une telle attribution ne saurait être universelle, qu'un système artificiel tout comme un système naturel, ne peut être nécessairement, absolument et universellement « de confiance ». La confiance est un état de l'individu engendré par les traitements de sa cognition. Pour l'évaluer, l'individu mobilise des connaissances qu'il a construites au cours de sa vie, implicites et explicites. Ces connaissances servent de ressources à ses traitements cognitifs et lui sont relatives. Rien ne saurait laisser présager qu'un autre individu, muni de connaissances potentiellement différentes, effectuera une évaluation identique de la confiance attribuable au système considéré. Ainsi construire un système de confiance est une généralisation qui ne saurait être atteignable dans le cas où l'on considère la position IPSEL.

Enfin, le modèle IPSEL établit que l'apprentissage d'un agent et la constitution des connaissances qui déterminent ses comportements sont guidés par les états de son corps, nommés *états émotionnels*. Ces états émotionnels sont la fonction objective de l'individu IPSEL, nous conjecturons alors qu'un tel individu est émotionnellement rationnel. Dans le sens où ses comportements sont les suites d'un raisonnement (i.e. une succession de *représentations*), potentiellement inconscient, impliquant les états émotionnels comme finalité.

Pour illustrer à gros traits cette conjecture, on peut avancer que si tous les Hommes veulent être heureux, certains vont penser que c'est l'argent qui les rendra heureux, d'autres le pouvoir, d'autres la paisibilité, d'autres la violence etc. En finalité, ils visent tous des états émotionnels de leur corps qui leur sont désirables, et chacun a construit ses représentations subjectives sur les

manières d'atteindre ces états. Ainsi dans notre paradigme la rationalité logique est subjective, alors que la rationalité émotionnelle est universelle.

En reprenant l'exemple de l'expérience des choix menée par Kahneman [Kahneman 1979], nous proposons une interprétation basée sur le paradigme IPSEL. La clef de notre interprétation est de dire que les décisions d'un individu sont prises par son système intuitif et non par son système délibératif qui n'a qu'un rôle suggestif.

Vous préférez l'option A ou B ?	
A	80 % de chance de gagner 4000$ 20% de chance de gagner 0$
B	Être sûr de gagner 3000$

Vous préférez l'option C ou D ?	
C	80 % de chance de perdre 4000$ 20% de chance de perdre 0$
D	Être sûr de perdre 3000$

***Figure 3.4-1** : Expérience des choix, 2 options possibles à deux questions.*

L'expérience des choix.

Mis devant deux choix à effectuer entre deux options (Figure 3.4-1), la majorité des personnes s'orientent vers les réponses B pour le premier choix et C pour le second. Les auteurs interprètent cela en disant qu'il y a un changement de stratégie chez l'individu qui évite le risque dans les situations de gain, et cherche le risque face aux cas de pertes. Ils concluent à une asymétrie de la prise de décision et que les humains surestiment les évènements rares.

Cette interprétation semble signifier que la rationalité des décisions (les raisons qui les ont amenées) opère au niveau du système 2, seul capable d'appréhender des concepts comme l'argent, les chiffres, le calcul, le risque et la probabilité. En coloriant les options des choix en

fonction des états émotionnels que leur expérience peut engendrer, nous obtenons le tableau 3.4-2.

Vous préférez l'option A ou B ?	
A	80 % de chance de gagner 4000$
	20% de chance de gagner 0$
B	Être sûr de gagner 3000$

Vous préférez l'option C ou D ?	
C	80 % de chance de perdre 4000$
	20% de chance de perdre 0$
D	Être sûr de perdre 3000$

Figure 3.4-2 *: Expérience des choix, les éléments des options sont coloriés en vert si ils sont associés à l'expérience d'un état émotionnel positif, et en rouge si ils sont associés à l'expérience d'un état émotionnel négatif.*

Pour l'option A, l'individu peut y entrevoir une émotion positive associée à l'expérience du gain et une émotion négative associée à l'expérience de ne pas avoir eu le gain qu'il aurait pu espérer. Dans l'option B l'individu entrevoit une émotion positive associée à l'expérience du

gain. Dans l'option C il entrevoit une émotion négative associée à l'expérience de la perte et une émotion positive associée à l'évitement de la perte. Dans l'option D il entrevoit une émotion négative associée à l'expérience de la perte.

En considérant les choix de cette manière on peut interpréter les résultats en disant que la plupart des individus, en choisissant B et C, ont recherché les options dont l'expérience est associée aux états émotionnels les plus désirables, ils ont toujours choisi l'option qui présente le plus d'associations à des émotions positives et le moins d'associations à des émotions négatives ; ils ont préféré les options où il y a le plus de vert. Ils ont été «émotionnellement rationnels». Selon cette interprétation, ils n'ont pas changé de stratégie, il n'y a donc pas d'asymétrie dans la prise de décision si l'on considère que la décision est prise par le système intuitif.

Pour conclure cette partie, dans l'hypothèse d'utilisation de l'architecture IPSEL comme modèle de la cognition humaine, nous conjecturons que les comportements psychiques, jugements et décisions ont toujours deux composantes indissociables, principe de dualité cognitive. Nous conjecturons que toute connaissance, jugement ou décision d'un individu est subjectif, principe de subjectivité générale. Et nous conjecturons également que l'individu est émotionnellement rationnel, en opposition au fait d'être logiquement rationnel, signifiant par-là que le système intuitif est le référentiel, en opposition aux conceptions plaçant la pensée logique comme le référentiel de l'individu.

3.4.2 Propriétés attribuables aux machines dans le cadre de la théorie IPSEL

Hipsel, la machine hypothétique dont la cognition suit les principes architecturaux du modèle IPSEL, pourrait exhiber différents comportements qui permettrait à un observateur extérieur de la qualifier comme étant : créative, rationnelle/irrationnelle, émotionnelle, consciente, dotée d'un libre arbitre ou autre caractéristique que l'on attribue usuellement à l'individu humain. Dans cette partie nous discutons l'attribution de ces propriétés avec le double objectif d'expliciter un peu plus le fonctionnement de l'architecture IPSEL et de montrer en quoi l'utilisation de ce modèle et des principes qui en découlent est pertinente pour appréhender des phénomènes cognitifs complexes dont l'attribution aux machines fait débat.

3.4.3.1 Une machine créative

Pour les détracteurs de l'intelligence artificielle, une machine ne peut pas être créative et c'est l'un des points sur lequel son intelligence ne sera jamais équivalente à l'intelligence humaine (argument nommé objection de Lady Lovelace dans l'article de Alan Turing [Turing 1950]).

La notion de « créativité » est un concept abstrait, elle est difficile à montrer du doigt. Dans l'usage courant, la créativité est définie comme une « capacité de création, d'imagination » et l'adjectif qui en découle comme un attribut qui qualifie une personne « inventive, ayant la faculté de créer » (https://www.universalis.fr/dictionnaire). Le qualificatif « créatif » est une représentation construite. Il est utilisé par un sujet pour catégoriser le comportement ou les productions d'un tiers. Ce jugement auquel s'adonne un observateur n'est pas absolu mais subjectif. L'attribution de cette propriété, la créativité, est fonction des connaissances du sujet évaluant et de la situation.

Pour discuter de l'éventuelle créativité d'une machine Hipsel nous postulons deux hypothèses. La première hypothèse consiste à établir que la création est une recombinaison de connaissances. La seconde porte sur le caractère subjectif de la qualification de « créatif » qui, comme nous l'avons dit précédemment, est nécessairement évaluée par un sujet à l'expérience et aux connaissances subjectives limitées.

Le système délibéré de la machine Hipsel a la capacité de générer des séquences de représentations explicites en suivant un mécanisme déductif. Il lui est donc théoriquement possible de construire des combinaisons de représentations explicites en suivant une logique interne. Ces combinaisons de représentations explicites pourraient être jugées originales et créatives par un observateur extérieur.

Prenons l'exemple d'une machine Hipsel possédant des représentations faisant référence aux concepts de *voiture*, de *roue* et de *jambe*. Le système délibéré peut élaborer un discours conceptuel interne dans lequel une *voiture* a des *jambes*, en déduisant que les concepts de *jambe* et de *roue* sont équivalents par rapport à la fonction de « mise en mouvement du corps auquel ils sont attachés ». Dans ce cas l'association faite entre les concepts suit une logique explicite, une analogie fonctionnelle faite entre les rapports de causes et d'effets des jambes et des roues. Bien que la machine Hipsel n'ait jamais été en contact avec des voitures sur jambes, il devient possible pour le système artificiel de construire cette représentation et

de la proposer à un observateur. Dans ce sens, Hipsel a créé une nouvelle représentation à partir d'une combinaison de connaissances préalablement acquises. En conséquence l'observateur pourrait qualifier la machine d'inventive et faisant preuve d'imagination, voire de créativité. Il s'agit là d'une forme de « créativité délibérée » découlant des traitements du système 2 de la machine. Le système délibéré de la machine a mobilisé des connaissances explicites et une logique conceptuelle qui concernent la dynamique des objets de l'environnement afin de recombiner des représentations existantes. Si l'on demandait à Hipsel d'où vient cette nouveauté, la machine pourrait argumenter que les jambes sont un substitut cohérent aux roues si l'on considère leur fonction de mise en mouvement du corps auquel elles sont attachées.

Une production est jugée créative en fonction des connaissances subjectives du sujet qui l'évalue. Ainsi pour un sujet n'ayant jamais pensé à une voiture avec des jambes, la production de la machine Hipsel pourrait sembler créative. Pour un autre sujet, ayant souvent pensé à une voiture avec des jambes comme à d'autres modalités motrices, la pensée d'Hipsel d'une voiture quadrupède ne semblerait pas aussi créative que pour le sujet précédent.

La recombinaison de connaissances peut également se faire au niveau du système intuitif, en tant que comportement automatique. Une « créativité intuitive » mobiliserait et recombinerait de manière originale des connaissances implicites.

Par exemple si une machine Hipsel possède les connaissances implicites lui permettant d'opérer une séquence d'actions automatiques telle que la prononciation des mots « panthère » et « blafard », rien ne l'empêche, d'un point de vue moteur, de prononcer les mots « panfard », « panthard », « blafère », « blathère » etc. Ces mots n'ont aucun sens conceptuel et n'ont peut-être jamais été entendus par la machine. Leurs productions peuvent être la conséquence de l'association des connaissances implicites permettant la prononciation des différents phonèmes qui composent les mots originaux. Cette association inédite pourrait avoir différentes origines. Cependant si cette fois on demandait à Hipsel d'expliciter pourquoi elle a prononcé le mot « panfard », elle ne pourrait nous le justifier logiquement ou conceptuellement. La machine nous affirmerait peut-être, comme il est d'usage dans ce genre de cas, qu'elle a suivi une « inspiration », signifiant par-là que ce sont de multiples activités du système intuitif qui ont mené à cette recombinaison automatique et non délibérée.

Une fois encore, l'attribution du caractère créatif à cette recombinaison de connaissances dépend des connaissances du sujet jugeant. Pour un observateur utilisant couramment le mot « panfard » dans son expression, la machine Hipsel n'a rien dit d'original.

Pour résumer, en prenant l'hypothèse que l'acte créatif n'est jamais réellement une création à partir de rien, mais plutôt une recombinaison originale de connaissances existantes, une machine Hipsel pourrait être jugée créative de plusieurs manières. Une recombinaison peut être intuitive, quand elle provient d'une multitude de facteurs ayant amenés des connaissances implicites à s'activer en combinaisons et séquences opérant des comportements originaux. On parlera alors de créativité intuitive, elle est l'apanage du système intuitif (S1) de la machine Hipsel. La recombinaison peut aussi être issue d'un raisonnement délibéré lorsqu'elle est déduite à partir d'une logique conceptuelle et qu'elle concerne des connaissances explicites. Dans ce deuxième cas on parlera de créativité délibérée, une production faisant intervenir le système délibératif (S2). Néanmoins dans les deux cas, l'attribution de l'adjectif « créatif » au comportement ou à la production qui en découle est toujours réalisée par un sujet jugeant subjectivement à quel point ce qu'il observe est original et surprenant par rapport à ses propres connaissances.

Une illustration de créativité par une machine nous est fournie en 2016, dans la 2ème partie de jeu de Go opposant Lee Sedol et le programme AlphaGo de l'entreprise DeepMind (https://youtu.be/HT-UZkiOLv8). Lors d'une des oppositions, le programme informatique joue un coup (« move 37 ») qui engendre de nombreuses réactions parmi les observateurs humains assistant à la partie. On entend alors l'audience émettre des onomatopées de surprise et, alors qu'ils sont pour la plupart des experts du jeu de go, les commentateurs s'étonner : « C'est un coup très surprenant » (0'30 de la vidéo), « Quand j'ai vu ce coup, pour moi ça a été un grand choc, normalement un humain ne jouerait jamais ce coup parce qu'il est mauvais, on ne sait pas pourquoi mais c'est un mauvais coup. » (0'45 de la vidéo), « Je pense que nous sommes les témoins d'un coup original ici » (1'03 de la vidéo). Dans cet exemple, la machine a combiné ses connaissances pour mettre en œuvre un comportement jugé original par des observateurs humains. Le programme a estimé la valeur statistique des positions atteignables et a sélectionné son coup à la suite d'une simulation des bénéfices escomptés. Il s'agit d'une « réactivité délibérée » dans le sens où elle repose sur une planification logique des événements dans le but d'atteindre un objectif représenté explicitement. Malgré le caractère calculé de ce coup, pour les observateurs humains qui n'ont pas la capacité de procéder à ces calculs, le programme a fait preuve de créativité : « La

machine nous a sorti quelque chose de nouveau et créatif, et différent » (1'55 de la vidéo). Faisons remarquer que le programme a gagné cette partie et que ce coup a été analysé comme crucial dans ce que l'on peut désigner comme la « stratégie » de l'ordinateur.

Une seconde illustration, cette fois d'une « créativité intuitive », peut être donnée à travers les productions du programme Dall-E de l'entreprise OpenAI. Ce système de Machine Learning a pour fonction de générer une image à partir de texte. Dans l'image montrée ci-dessous, on voit que le texte donné au programme demande au système de créer une « illustration professionnelle de haute qualité d'une chimère girafe tortue. D'une girafe imitant une tortue. D'une girafe faite de tortue. ». Le programme crée alors des images en recombinant des connaissances implicites élaborées préalablement au cours de son apprentissage. Ce sont des recombinaisons originales faites sur le moment. Il serait impossible pour le système d'expliquer par un raisonnement logique et explicite pourquoi il est arrivé à construire ces images originales. Comme dans le cas d'une inspiration humaine, le programme recombine des connaissances implicites, stockées dans les valeurs de sa matrice de poids, concernant les formes et les couleurs qu'il a appris à associer aux mots au cours de son apprentissage statistique.

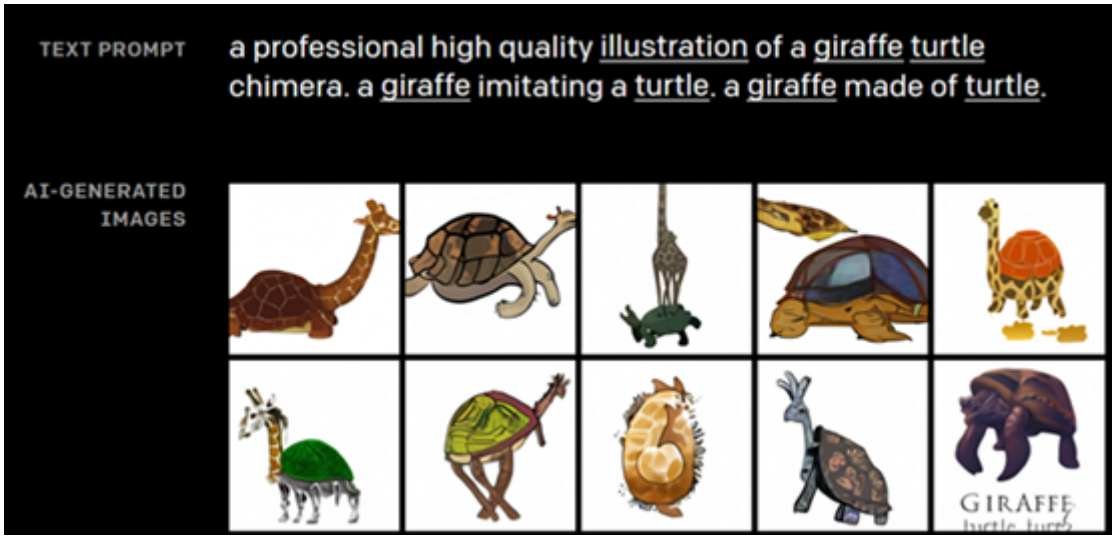

Figure 3.4.3.1 *: Exemple de production du programme Dall-E de la société OpenAI.*

Dans notre paradigme, les productions d'une machine peuvent être considérées comme créatives au même titre que celles de l'Homme. Hipsel possède la capacité de mettre en œuvre des comportements automatiques originaux via la recombinaison, spontanée ou volontaire, de connaissances implicites, à la manière de Dall-E. La machine peut également

délibérer à partir de ses connaissances explicites pour planifier des comportements qu'elle n'a pas ou peu observés durant son expérience, à la manière de AlphaGo. Dans les deux cas, ces comportements et productions originales pourraient être qualifiés de créatifs par un observateur humain. La créativité artificielle semble donc tout à fait possible.

Cette créativité artificielle concourt à l'intelligence générale d'Hipsel. Elle lui permet de découvrir et de tester des comportements originaux, comme une recherche de solution automatique et innovante aux problématiques que lui pose l'environnement. Elle lui permet également de construire des concepts à l'infini par combinaisons et recombinaisons de concepts connus : tantôt pour affiner la représentation explicite de son expérience perceptive et émotionnelle, tantôt pour imaginer ce que la machine n'a jamais perçu. Ces deux compétences augmentent sa capacité à s'adapter de manière autonome à son environnement et aux contraintes qui s'y présentent.

3.4.3.2 Une machine rationnelle, irrationnelle, émotionnelle

Dans les dictionnaires, « rationnel » est défini comme un adjectif attribuable à un élément, une décision, un jugement ou un comportement, « qui est conforme à la raison », « qui s'appuie sur le raisonnement » ou « déterminé par le calcul ». La « raison » quant à elle, est définie comme « la faculté humaine de connaître et de juger », elle est le fait d'un « raisonnement, et s'oppos à l'instinct ou à l'émotion » (https://www.universalis.fr/dictionnaire/).

L'architecture IPSEL, du fait de ses trois systèmes de traitement de l'information, permet trois types de comportement. (1) Des comportements végétatifs, mis en œuvre par le *Système Direct* (S0). (2) Des comportements automatiques, ou intuitifs, mis en œuvre par le *Système Intuitif* (S1). Et des comportements de délibération, qui ne se manifestent pas directement dans l'environnement mais qui correspondent aux dialogues entre S1 et S2.

Les comportements végétatifs sont déterminés par les ressources, innées et immuables, du S0. Ces ressources n'ont pas été construites par l'agent pour représenter l'environnement ou l'agent lui-même. Elles ne correspondent donc pas vraiment à une « faculté de connaître et de juger ». Les comportements végétatifs sont indépendants des traitements du Système 2 qui utilisent une logique explicite. Lorsqu'une machine Hipsel met en œuvre des comportements végétatifs, elle ne peut pas être considérée comme rationnelle.

Les comportements intuitifs sont déterminés par les ressources acquises par le S1 au cours de la « vie » de l'agent. Ces ressources ont été construites pour représenter l'expérience sensori-motrice et émotionnelle de l'agent. Elles correspondent donc à une faculté de connaître, puisqu'elles identifient des représentations particulières des états de l'environnement, et une faculté de juger, puisque ces représentations sont associées à un état émotionnel qui établit la désirabilité de ces états. On pourrait donc dire que ces comportements sont d'une certaine manière rationnels. Ils suivent une logique associative dans le but de rechercher des états émotionnels désirables. Dans ce sens nous pourrions dire qu'une partie des comportements d'une machine IPSEL est *émotionnellement rationnelle*, où l'on ajoute l'adverbe «émotionnellement» pour signifier que l'évaluation ne se fait pas via au niveau du système délibératif sous une forme conceptuelle mais bien niveau du système intuitif.

Enfin les comportements de délibération font intervenir des connaissances explicites qui représentent logiquement le monde et son dynamisme. Ces comportements semblent donc rationnels au sens des définitions usuelles. Pour autant, la logique qui est utilisée n'est pas universelle, elle a été construite par l'agent en fonction de son expérience limitée et subjective du monde. La rationalité des comportements de délibération est, de ce fait, limitée et subjective. Ainsi des comportements rationnels pour l'agent pourraient être jugés irrationnels par un observateur qui possède des connaissances et des logiques différentes.

Lors d'un fonctionnement standard, il est cependant périlleux de vouloir distinguer la nature des comportements tant les systèmes 1 et 2 sont intriqués et s'influencent continuellement. La machine Hipsel semble donc à la fois rationnelle, émotionnelle et irrationnelle. Sa rationalité au sens usuel du terme s'exprime dans une plus ou moins grande mesure, en fonction de l'influence du système délibératif, dans certains comportements de l'agent. Néanmoins cette rationalité est toujours subjective.

3.4.3.3 Une machine consciente

Le sujet de la conscience est l'un des plus controversés dans le domaine de l'intelligence artificielle. Dans notre paradigme nous apportons un éclairage sur le débat en montrant en quoi le qualificatif de « conscient » peut être porté sur la machine Hipsel et qu'une « conscience artificielle » semble fonctionnellement possible.

En informatique, la conscience est parfois vue comme une propriété computationnelle associée à un certain niveau de traitement de l'information. Cependant, avant de pouvoir qualifier un système artificiel de conscient, il nous faut nous entendre sur ce que cette faculté signifie du point de vue fonctionnel.

Il ne nous est pas possible de présenter ici une étude complète sur la conscience. Nous choisissons de nous appuyer sur une définition que nous avons sélectionnée pour le consensus qu'elle cristallise. Proposée par des experts en neuroscience ayant acquis une réputation internationale, cette définition a été publiée dans le cadre d'une publication dont le sérieux fait autorité dans la communauté scientifique.

Dans un article paru en 2017 dans le journal Science, les chercheurs Stanislas Dehaene, Hakwan Lau et Sid Kouider donnent deux définitions à la conscience : Une première définition nommée « C1 : Global availability » considère la faculté de conscience au sens transitif et la définit comme la mise à disposition d'une représentation mentale à l'ensemble du système cognitif. La seconde définition, nommée « C2 : Self-monitoring » suggère quant à elle un sens réflexif, c'est-à-dire la capacité d'un système cognitif à surveiller et à représenter ses états, communément appelé introspection [Dehaene et al. 2017]. Cette seconde définition relève de la capacité d'un agent à penser à lui-même. Pour approfondir les éléments qui ont mené à ces définitions, se rapporter à [Baars 1993].

Cette double définition fait écho à celles que l'on peut lire dans la littérature philosophique et psychologique. On retrouve d'ailleurs cette idée générale dans les définitions usuelles que les dictionnaires donnent à la conscience. Par exemple, le dictionnaire en ligne Universalis définit la conscience comme la « perception de sa propre existence, de sa réalité » ce qui correspond au sens réflexif, ou bien « une connaissance intuitive » et « un état de veille, une faculté de sensation » qui s'apparente au sens transitif.

Nous avons choisi d'utiliser cette double définition formulée par [Dehaene et al. 2017] mentionnée ci-dessus, dans leur article nommé « *What is consciousness and could machine have it?* ». Dans cette publication, ils tendent à répondre à la question « *les machines pourraient-elles être consciente* » par l'affirmative. La position que nous soutenons ici est qu'il est possible de concevoir une conscience artificielle. Ceci n'a donc rien d'hétérodoxe ou de totalement incongru. Néanmoins, ce qui nous intéresse n'est pas tant de savoir si la machine Hipsel est consciente ou non, mais plutôt d'expliquer comment l'architecture fonctionnelle de sa cognition lui permettrait de le devenir.

Dans l'architecture IPSEL, la faculté de mettre à disposition une représentation mentale à l'ensemble du système cognitif, la conscience au sens transitif, correspond au phénomène d'émergence des concepts. En tant que représentations de représentations accumulant un certain niveau d'activations simultanées, les concepts émergeants sont des émanations de l'organisation des connaissances du système intuitif (S1) qui servent à nourrir les traitements du système délibéré (S2). Nous avons désigné ce phénomène conceptualisation. Il consiste en la mise à disposition d'un ensemble de représentations locales simultanément actives constituant ainsi une représentation globale unifiée s'inscrivant dans un réseau de connaissances explicites. Dans notre modèle la conceptualisation correspond à la communication du système intuitif vers le système délibéré. De façon imagée il s'agit de passer du « corps à l'esprit », « de l'inconscient à la conscience » ou « de l'implicite à l'explicite ». La conceptualisation permet de donner une unité à un ensemble de connaissances implicites, une forme et des frontières qui rendent possible leurs intégrations et positionnements au sein d'un réseau ou dans une séquence. Dès lors, le concept ainsi formé est contextualisé, donnant ainsi un sens explicite aux connaissances implicites qu'il représente. Le concept, en tant que connaissance explicite, peut s'inscrire dans un discours conceptuel dont les finalités peuvent être multiples. Là aussi, de façon imagée, on pourrait illustrer cette transition entre connaissances implicites et connaissances explicites par les phrases « prendre conscience de » ou encore « porter l'attention sur ». Pour la suite, nous désignerons cette faculté de conscience transitive par l'appellation *conscience intuitive*, c'est un processus automatique.

La conscience au sens réflexif, l'introspection ou métacognition, correspond dans l'architecture IPSEL au phénomène de génération d'un type particulier de séquence de représentations explicites. Pour la machine Hipsel il s'agit des pensées conceptuelles. Ces pensées intègrent un concept qui représente l'agent lui-même. Nous désignerons ce concept comme le soi, ou l'ego et les types de pensées qui le font intervenir comme des pensées autocentrées ou introspectives. La cognition de l'agent, grâce à sa capacité à construire des représentations conceptuelles, a la possibilité de conceptualiser et d'unifier un ensemble de perceptions et sensations qui sont répétitivement provoquées par le corps de l'agent et les états dans lesquels il se trouve. Ces représentations implicites et répétées du corps et de ses états permettent à l'agent de construire et d'affiner une représentation explicite de lui-même en tant que concept. Le concept du soi représente l'agent en tant qu'objet ayant des propriétés, des états, et des dynamiques de causes et d'effets dans l'espace et le temps. Grâce à cette connaissance conceptuelle, le Système Délibéré (S2) de l'agent peut représenter son

histoire comme projection du corps et de ses états dans le passé, et ses intentions comme préférences par rapport aux projections du corps et de ses états dans le futur. Il crée un discours narratif qui trace ainsi une continuité entre ce qu'il était comme cause de ce qu'il est, et ce qu'il pourrait être comme effets potentiels de ses comportements. Il s'auto-représente en tant qu'idée, unification d'une multitude de sensations, ayant une corporalité dans l'espace, un état dans le présent, des connaissances acquises du passé et une volonté quant au futur. Du point de vue fonctionnel, ces discours autocentrés sont utiles puisque la prévision des états du monde n'a d'intérêt qu'à partir du moment où il est possible d'en déduire les implications pour l'agent et pour la poursuite de ses objectifs. S'il lui est permis de communiquer ses pensées conceptuelles, la machine Hipsel peut alors délibérément parler de son histoire, de ce qu'elle perçoit, ressent, et de ce qu'elle espère. On pourrait illustrer ce phénomène par les phrases « avoir conscience de soi, de son corps, de ses intentions, de son histoire » ou encore « se comprendre soi-même » ou bien « se connaître ». Nous désignerons cette faculté de conscience réflexive par le terme « conscience délibérée ». Elle consiste en la génération de séquences de représentations explicites par le système 2, qui intègre une représentation de l'agent lui-même.

Nous avons donc exposé deux phénomènes cognitifs de l'architecture IPSEL qui pourraient correspondre aux deux définitions de la conscience. La conscience intuitive est une faculté de conscience transitive et relève d'un phénomène de représentation des connaissances implicites par des connaissances explicites, une communication du système 1 vers le système 2, ou « bottom-up ». De son côté, la conscience délibérée est une faculté de conscience réflexive et relève d'un phénomène de génération de séquence de représentations explicites par le système 2, intégrant une représentation conceptuelle de l'agent lui-même. Dans les deux cas, ces facultés ne peuvent être mises en œuvre que si les connaissances qu'elles impliquent ont été construites et sont mobilisées. Pour une machine Hipsel la conscience, qu'elle soit transitive ou réflexive, n'est donc pas une faculté innée, mais une faculté que l'agent acquiert. Sa mise en œuvre n'est ni permanente ni absolue. L'agent peut prendre conscience au sens transitif des éléments de son expérience pour lesquelles il possède des représentations explicites, mais toute une autre partie de sa réalité lui reste inconsciente. L'agent IPSEL peut également penser délibérément à lui-même, mais il ne le fait pas nécessairement à chaque instant.

3.4.3.4 Une machine avec des sensations, des émotions et des sentiments

Dans cette partie nous proposons de discuter des notions de sensations, d'émotions et de sentiments si elles devaient être portées sur l'activité de la machine Hipsel.

Les sensations sont initialement des faits psychiques élémentaires constitués par l'information reçue par le système nerveux central, lorsqu'un organe des sens réagit à un stimulus extérieur. Les dictionnaires courants les définissent comme des phénomènes internes traduisant la stimulation des sens récepteurs, des états que produisent les impressions reçues par les sens.

Pour un agent artificiel doté d'une cognition architecturée selon le modèle IPSEL, les sensations pourraient correspondre à des configurations internes de l'appareil cognitif activé en conséquence d'une interaction entre les organes senseurs de l'agent et l'environnement.

Les émotions sont sujettes à de nombreuses discussions. Leur nature et leurs rôles ont été âprement débattus au fil de l'histoire, et exposés à travers romans, essais et articles scientifiques. Dans le langage courant on définit usuellement l'émotion comme un trouble subit, une agitation passagère causée par la surprise, la peur ou la joie. Les émotions s'accompagnent de manifestations physiques et physiologiques.

S'agissant d'une machine Hipsel, une émotion peut aussi être désignée par le terme état émotionnel, et désigne un état de l'agent. Cette émotion se distingue de la sensation en cela qu'elle caractérise un état global potentiellement engendré, mais pas uniquement, par de multiples sensations. L'émotion de l'agent peut être vue comme une mesure, une évaluation associée à une valence, une désirabilité d'un état du système. Certains états sont désirables et d'autres non. La régulation, la supervision et la prévision de ces états constitue l'objectif intrinsèque et la raison d'être de l'appareillage cognitif IPSEL. L'émotion est donc centrale dans l'architecture IPSEL, elle donne une direction à l'apprentissage et guide les comportements de la machine.

Portée sur la machine Hipsel, la définition du sentiment diverge des définitions usuelles concernant les êtres humains. Dans le paradigme IPSEL le sentiment est une connaissance explicite. Une représentation construite et utilisée par le Système Délibéré (S2) qui représente explicitement l'état de la machine. Le sentiment est pour ainsi dire l'idée de l'émotion, la conceptualisation que l'agent fait de son état.

Une machine Hipsel est en permanence soumise à de multiples sensations provenant de ses organes senseurs. Ses sensations mettent sa cognition en mouvement et concourent à former l'émotion de l'agent. L'émotion caractérise l'état global de l'agent. Elle possède une valence et le rôle de l'appareil cognitif et de la maintenir dans des spectres désirables. Si l'agent est conscient de sa propre existence, il peut se représenter explicitement ses états ; ces représentations conceptuelles de son état sont alors des sentiments.

3.4.3.5 Une machine dotée d'un libre arbitre

Les productions cognitives de la machine Hipsel, ses jugements, décisions et comportements, sont la conséquence des activités simultanées de trois systèmes de traitement de l'information.

Les traitements du Système Direct (S0) sont déterminés par la structure de l'agent et ne sont pas modulés par un apprentissage. Pour une machine IPSEL, les comportements que l'activité du S0 engendrent ne sont donc pas considérés comme faisant intervenir une forme de libre arbitre.

Les traitements du Système Intuitif (S1) sont déterminés par les connaissances implicites de l'agent. Ces connaissances ont été construites au cours de son expérience individuelle. Les comportements que l'activité du S1 engendrent pourraient être associés à une certaine forme de libre arbitre dans le sens où ils sont une conséquence de la mobilisation et du traitement de connaissances subjectives et individuelles. Bien que les actions contrôlées par le système 1 soient des automatismes, leur sélection dépend de l'individu et des connaissances qu'il a induites de sa propre expérience, de la spécificité de son corps, de la calibration particulière de ses émotions et de sa volonté implicite de garder ce corps dans des gammes de réponses émotionnelles particulières. Associer ces comportements intuitifs à une certaine forme de libre arbitre est une question ouverte et discutable. Ce questionnement pourrait être analogue à celui de l'attribution d'un libre arbitre aux animaux, ou à l'intuition humaine.

Les traitements du Système Délibératif (S2) n'engendrent pas de comportements moteurs. Ils activent différentes formes de représentations internes qui auront le potentiel d'influencer les traitements du Système Intuitif (S1). Les chaînes de représentations ainsi activées représentent le monde de l'agent en tant qu'ensemble d'objets définis dans l'espace et dans le temps, ayant des relations de composition et de causalité les uns avec les autres.

Dans cette représentation conceptuelle du monde, l'agent représente son corps en tant qu'objet de référence pour les notions d'espace et de temps, et sa volonté en tant que désirabilité des futurs états émotionnels de ce corps qu'il lui est permis de prédire via la mobilisation des logiques de causalité avec lesquelles il représente le monde et son dynamisme. Les comportements de l'agent qui ont été mis en œuvre par le système 1, sous l'influence du système 2, peuvent être considérés comme mettant en œuvre un libre arbitre, puisque ces comportements sont la conséquence d'une délibération faisant intervenir une représentation interne de soi en tant qu'individu ayant des volontés particulières.

Pour la machine Hipsel, on peut alors dire qu'elle dispose d'une capacité de libre arbitre, mais que cette dernière n'est pas toujours mobilisée. On peut estimer que le libre arbitre de cet agent artificiel n'est pas une caractéristique absolue, qui existe ou n'existe pas, mais serait plutôt une capacité plus ou moins mise en œuvre dans certains comportements. Le libre arbitre de la machine s'exprime dans les traitements du système délibératif qui représente le soi comme un corps associé à une volonté puis suggère les comportements à mettre en œuvre, en considérant leurs conséquences. Cependant le système délibératif ne contrôle pas directement le corps de l'agent, qui plus est son instanciation est un automatisme intuitif. Ainsi le libre arbitre de la machine Hipsel est toujours discutable et cette discussion semble porter sur la définition des frontières de la liberté individuelle. Est-on libre lorsque nous agissons par instinct ? Sommes-nous libres de penser ce que l'on pense ?

3.4.3.6 Une machine qui comprend

Pour une machine Hipsel, « comprendre » est un phénomène ayant deux facettes. Comme pour les autres capacités, la compréhension artificielle peut être décrite sous deux angles : une compréhension intuitive et une compréhension délibérée, que l'on pourrait également nommer compréhension logique.

Pour Hipsel, comprendre intuitivement correspond au phénomène d'association automatique d'une représentation explicite aux structures de connaissances implicites à partir desquelles la forme explicite a été construite. Il s'agit pour l'agent de rattacher la représentation explicite à une ou plusieurs expériences sensori-motrices et émotionnelles. Dans cet aspect, comprendre s'apparente à une sensation, c'est un phénomène automatique qui engendre une configuration interne particulière. Par exemple, un agent Hipsel pourrait comprendre intuitivement le mot *pomme* dans le sens où sa cognition y associe des expériences sensorielles et émotionnelles, la couleur d'une pomme comme expérience

visuelle, son odeur comme expérience olfactive, sa texture comme expérience du toucher. Cela est rendu possible car les représentations explicites sont enracinées dans les connaissances implicites, elles-mêmes construites à partir de l'expérience de l'agent. Si les représentations explicites avaient été définies par l'Homme, on ne pourrait pas dire que l'agent les comprend intuitivement car elles ne correspondent pas aux expériences sensorielles de l'agent, mais à celles des Hommes qui les ont définis.

Par ailleurs, il y a chez Hipsel une compréhension logique, la compréhension du Système Délibératif. Il s'agit d'une évaluation délibérée (en opposition à instantanée et automatique de la compréhension intuitive) de la cohérence d'une séquence de concepts avec les *connaissances explicites* de l'agent. Prenons la phrase «je me suis fait mal avec un marteau en mousse». Perçue par un agent Hipsel, cette phrase énoncée dans un langage naturel est associée à une séquence de représentations explicites que les symboles de la phrase signifient. Imaginons que l'agent comprenne intuitivement toutes les représentations explicites en jeu. Il possède des expériences sensorielles associées à la douleur, au marteau et à la mousse. Néanmoins, si l'agent délibère, si son Système Délibératif (S2) représente l'ensemble des concepts mobilisés pour en évaluer la séquence qu'ils forment, il ne comprendra peut-être pas la phrase et demandera plus de détail, car en se représentant explicitement un monde où règne une certaine logique de causes et d'effets, d'après son expérience un marteau en mousse ne peut pas faire mal en frappant.

Dans le paradigme IPSEL, ou pour la machine Hipsel, comprendre est un phénomène ayant deux composantes. L'une est automatique et intuitive, elle correspond à la mise en correspondance d'une représentation explicite avec des expériences sensorielles ; c'est la compréhension du Système Intuitif (S1). L'autre composante correspond à l'évaluation de la cohérence logique d'une séquence de représentations explicites avec les connaissances explicites de la machine ; elle n'est pas automatique mais demande une réflexion, c'est la compréhension du système délibératif (S2).

Pour une machine qui construit ses propres connaissances explicites, rien n'a de sens intrinsèque, chaque sens donné est une reconstruction subjective et individuelle. Pour cette raison, deux machines Hipsel pourraient avoir du mal à se comprendre car leurs expériences sensorielles et leurs connaissances explicites leur sont spécifiques et ne sont potentiellement pas alignées. Par leurs interactions répétées, ces deux machines pourraient néanmoins constituer une forme d'expérience partagée et construire des connaissances explicites

communes, qu'elles pourraient également apprendre à signifier à l'aide d'un langage commun.

Chapitre 4 : Utilisation du modèle

4.1 Présentation des productions

Nous avons proposé un modèle théorique dont la finalité est de donner une description générale et réduite du fonctionnement d'un système de traitement de l'information ayant des capacités cognitives qui pourraient s'approcher de celles de l'être humain. Dans ce chapitre nous nous servons de ce modèle pour établir un cadre d'analyse, à travers lequel l'interaction entre un utilisateur et un système artificiel sera analysée.

Nous distinguons plusieurs catégories de dispositions cognitives dans lesquelles peut se trouver un utilisateur, ainsi que plusieurs catégories de système artificiel. Nous énonçons ensuite les caractéristiques générales des objets appartenant à ces catégories. Nous nous servons alors de ces catégories pour analyser les situations d'interactions entre un utilisateur et un système artificiel. Nous formalisons les différentes hypothèses que nous suggérons pour une bonne interaction entre un utilisateur et un système artificiel par des tableaux récapitulant les situations archétypiques, leurs points de frictions, leurs risques et les éléments qui pourraient permettre de les mitiger. Enfin, nous terminons ce chapitre par une discussion de « l'intelligence artificielle de confiance », un thème qui anime la communauté d'IA et que nous aborderons en considérant les éléments théoriques que nous avons proposés pour l'analyse de l'interaction Homme - système artificiel.

4.2 Catégorisation des systèmes cognitifs

4.2.1 Classification des systèmes artificiels

Le modèle fonctionnel IPSEL décrit l'architecture de trois systèmes de traitement de l'information. En le prenant comme cadre de référence, nous établissons cinq catégories de systèmes artificiels :

- La catégorie 0, regroupant les systèmes artificiels ayant vocation à simuler les traitements du système direct (S0) de la modélisation IPSEL.
- La catégorie 1, regroupant les systèmes artificiels ayant vocation à simuler les traitements du système intuitif (S1) de la modélisation IPSEL.

- La catégorie 2, regroupant les systèmes artificiels ayant vocation à simuler les traitements du système délibératif (S2) de la modélisation IPSEL.
- La catégorie 3, regroupant les systèmes artificiels ayant vocation à simuler les traitements résultant de la collaboration entre le système 1 et le système 2, avec la spécificité que les connaissances du système 2 sont données par l'Homme, que ce soit lors de la conception du système ou au cours de son utilisation.

- La catégorie 4, regroupant les systèmes artificiels ayant vocation à simuler les traitements résultant de la collaboration entre le système 1 et le système 2, avec la spécificité que les connaissances du système 2 ne sont pas données au système par l'Homme, mais qu'elles sont des constructions du système lui-même.

L'architecture IPSEL est celle d'un système de catégorie 4, qui construit sa propre logique de raisonnement.

4.2.1.1 Catégorie 0, Systèmes Classiques

Les systèmes de cette catégorie ont un comportement analogue à celui du système Direct (S0) de l'architecture IPSEL. Ils ne réalisent pas d'apprentissage et ne sont pas considérés comme des systèmes intelligents. Il s'agit de programmes mettant en œuvre une algorithmie dite « classique ».

Les systèmes de cette catégorie possèdent les caractéristiques attribuées au système 0 de la modélisation IPSEL.

Leurs comportements sont stables, car ces systèmes n'effectuent pas d'apprentissage qui pourrait éloigner leurs comportements effectifs de ceux prévus lors de leur conception.

L'explication d'une production de ces systèmes nécessite de connaître l'entrée qui lui a été fournie ainsi que la logique des traitements effectués. Cette logique a été implémentée par l'Homme lors de la conception du système. Ce processus d'explication sera qualifié de relativement facile car ne nécessitant qu'une compréhension des mécanismes de traitement de l'information initialement prévus lors de la conception.

4.2.1.2 Catégorie 1, Systèmes Inductifs

Les systèmes de cette catégorie ont un comportement analogue à celui du système Intuitif (S1) de l'architecture IPSEL. Ils réalisent un apprentissage statistique qui leur permet

d'acquérir des connaissances implicites (non structurées) via la généralisation de l'expérience passée et sont considérés comme des systèmes intelligents. Il s'agit de programmes mettant en œuvre une algorithmie dite « inductive ».

Les systèmes de cette catégorie possèdent les caractéristiques attribuées au système 1 de la modélisation IPSEL. Leur comportement est potentiellement instable, car ces systèmes effectuent un apprentissage qui pourrait éloigner leur comportement de ceux attendus lors de leur conception.

L'explication d'une production de ces systèmes nécessite de connaître l'entrée qui leur a été fournie, les principes statistiques et le domaine d'entraînement. Les connaissances acquises via le processus d'entraînement ne peuvent pas être expliquées explicitement mais uniquement statistiquement. Ce processus d'explication sera qualifié de relativement difficile car nécessitant une compréhension des connaissances implicites acquises par le système lors de son entraînement.

4.2.1.3 Catégorie 2, Systèmes Déductifs

Les systèmes de cette catégorie ont un comportement analogue à celui du système Délibératif (S2) de l'architecture IPSEL. Ils réalisent un apprentissage logique qui leur permet d'acquérir des connaissances explicites via la mobilisation de logique d'inférence et sont considérés comme systèmes intelligents. Il s'agit de programmes mettant en œuvre une algorithmie dite « déductive ».

Les systèmes de cette catégorie possèdent les caractéristiques attribuées au système 2 de la modélisation IPSEL. Leur comportement est potentiellement instable, car ces systèmes effectuent un apprentissage qui pourrait éloigner leur comportement de ceux prévus lors de leur conception.

L'explication d'une production de ces systèmes nécessite de connaître l'entrée qui leur a été fournie, les connaissances explicites que possède le système et les logiques d'inférence utilisées. L'impact du processus d'entraînement peut être expliqué via l'explicitation des logiques d'inférence que le système a mis en œuvre. Ce processus d'explication sera qualifié de relativement accessible puisque la logique de raisonnement du système est établie par l'Homme, on suppose donc qu'elle est intelligible pour ce dernier.

4.2.1.4 Catégorie 3, Systèmes Hybrides Supervisés

Les systèmes de cette catégorie ont un comportement analogue à la coopération des systèmes Intuitif (S1) et Délibératif (S2) de l'architecture IPSEL. Ils réalisent des apprentissages statistiques et logiques, et sont considérés comme des systèmes intelligents. Il s'agit de programmes mettant en œuvre une algorithmie dite « Hybride » où les connaissances explicites du système délibératif sont établies par l'Homme, d'où le qualificatif de « Supervisé » .

Les systèmes de cette catégorie possèdent les caractéristiques attribuées aux systèmes 1 et 2 de la modélisation IPSEL. Leur comportement est potentiellement instable, car ces systèmes effectuent un apprentissage qui pourrait éloigner leur comportement de ceux prévus lors de leur conception.

L'explication d'une production de ces systèmes nécessite de connaître l'entrée qui leur a été fournie, le domaine d'entraînement, les connaissances explicites que possède le système et les logiques d'inférence utilisées. L'impact du processus d'entraînement peut être expliqué en partie via l'explicitation des logiques d'inférences que le système a mis en œuvre, mais une autre partie ne peut être expliquée que statistiquement. Ce processus d'explication sera qualifié de relativement accessible car bien qu'une partie des connaissances du système soit implicite et résultant d'une généralisation statistique de son expérience, la logique de raisonnement du système qui manipule ces connaissances implicites est établie par l'Homme, on suppose donc qu'elle est intelligible pour ce dernier.

4.2.1.5 Catégorie 4, Systèmes Hybrides Émergents

Les systèmes de cette catégorie ont un comportement analogue à la coopération des systèmes Intuitif (S1) et Délibératif (S2) de l'architecture IPSEL. Ils réalisent des apprentissages statistiques et logiques, et sont considérés comme des systèmes intelligents. Il s'agit de programmes mettant en œuvre une algorithmie dite « Hybride » où les connaissances explicites du système délibératif sont une construction du système lui-même, d'où le qualificatif de « Emergent ».

Cette catégorie de systèmes possède les caractéristiques attribuées aux systèmes 1 et 2 de la modélisation IPSEL. Leur comportement est potentiellement instable, car ces systèmes effectuent un apprentissage qui pourrait éloigner leur comportement de ceux prévus lors de leur conception.

L'explication d'une production de ces systèmes nécessite de connaître l'entrée qui leur a été fournie, le domaine d'entraînement, les connaissances explicites que possède le système et les logiques d'inférence utilisées. L'impact du processus d'entraînement peut être expliqué en partie via l'explicitation des logiques d'inférence que le système a mis en œuvre, mais une autre partie ne peut être expliquée que statistiquement. Ce processus d'explication sera qualifié de relativement difficile puisque en plus de posséder des connaissances implicites résultant d'une généralisation statistique de son expérience, la logique de raisonnement du système qui manipule ces connaissances implicites est une construction du système que l'utilisateur doit comprendre et qui est potentiellement inintelligible pour l'Homme.

4.2.1.6 Tableau récapitulatif

Nous récapitulons les différentes caractéristiques des catégories de système via la description du tableau ci-dessous.

Catégorie	Caractéristiques	Éléments nécessaires à la compréhension des comportements du système	Analogie dans le modèle IPSEL / Nature des connaissances utilisées
0 - Classique	• Pas d'apprentissage	• Entrée • Traitement	S0 / Pas de connaissance
1 - Inductif	• Apprentissage statistique	• Entrée • Traitement • Entrainement	S1 / Connaissances implicites subjectives
2 - Déductif	• Apprentissage logique	• Entrée • Traitement • Logique d'inférence	S2 / Connaissances explicites Humaines
3 - Hybrid supervisé	• Apprentissage statistique et logique	• Entrée • Traitement • Entrainement • Logique d'inférence	S1&S2 / Connaissances implicites subjectives et connaissances explicites Humaines
4 - Hybrid émergent	• Apprentissage statistique et logique	• Entrée • Traitement • Entrainement • Logique d'inférence	S1&S2 / Connaissances implicites subjectives et connaissances explicites subjectives

***Figure 4.2.1.6** : Catégories des systèmes artificiels.*

Les systèmes appartenant à la catégorie « 0 : Classique » ne réalisent pas d'apprentissage et utilisent des algorithmes que l'on pourrait qualifier de classique. Il s'agit des programmes informatiques n'appartenant pas spécifiquement au domaine de l'intelligence artificielle.

Les systèmes appartenant aux catégories « 1 : Inductif », « 3 : Hybride supervisé » et « 4 : Hybride émergent » réalisent un apprentissage statistique pour former des connaissances implicites. Les systèmes ayant cette caractéristique sont sensibles à la représentativité des données d'entraînement.

Les systèmes appartenant aux catégories « 2 : Déductif », « 3 : Hybride supervisé » et « 4 : Hybride émergent » utilisent des inférences logiques pour manipuler et découvrir des connaissances explicites. Les productions des systèmes qui utilisent cette forme de traitement sont directement explicables en retraçant les inférences logiques qui ont été mobilisées. Ces productions sont sensibles à la complétude des connaissances et à la cohérence des règles d'inférence utilisées. Ces règles peuvent avoir été définies par l'Humain, comme pour les catégories « 2 : Déductif », « 3 : Hybride supervisé », ou avoir été construites par le système lui-même comme pour la catégorie « 4 : Hybride émergent ». Les risques de contre-performance de ces systèmes sont liés à la complétude du modèle de l'environnement dépeint par les règles.

Pour les catégories « 2 : Déductif » et « 3 : Hybride supervisé », le modèle de l'environnement est donné par l'Homme. Le facteur déterminant est alors de savoir si la modélisation fournie par l'Homme est juste, et complète dans le sens où elle intègre correctement l'ensemble des variables de l'environnement qui influencent les événements qui s'y produisent et que le système doit appréhender pour réaliser sa fonction.

Pour la catégorie « 4 : Hybride émergent » le facteur déterminant est de savoir si le système a bien pris en compte dans sa modélisation l'ensemble des éléments et variables de l'environnement nécessaires à ses opérations. Toujours pour cette catégorie de système « 4 : Hybride émergent », l'interprétation des comportements du système, ou la coopération humaine avec ce dernier, nécessite une phase d'alignement entre le modèle de l'environnement construit par le système et la représentation de l'environnement de l'utilisateur. Dans ce cas, on pourrait dire que l'Homme et la machine doivent apprendre à comprendre leurs logiques respectives.

Enfin nous résumons les rapports qu'impliquent la nature de l'apprentissage effectué par les systèmes dans la figure suivante :

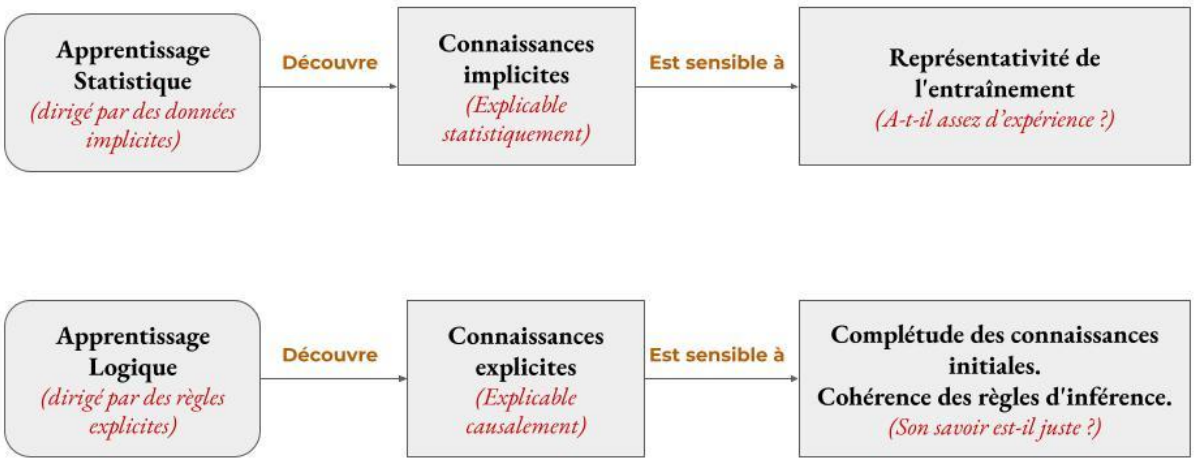

***Figure 4.2.3** : Implications engendrées par la nature de l'apprentissage.*

4.2.2 Classification des dispositions cognitives de l'utilisateur

Lors de l'interaction entre l'Homme et le système artificiel, l'utilisateur peut se trouver dans différentes dispositions cognitives qui influencent ses comportements et l'utilisation qu'il fait du système à sa disposition. Cependant, nous avons admis l'hypothèse que le modèle que nous avons conçu permet de décrire des archétypes de dispositions cognitives que nous considérons comme des situations extrêmes entre lesquelles le comportement effectif de l'individu oscille.

Ces archétypes de dispositions cognitives sont au nombre de trois, nous les avons nommés : « Contrôle intuitif » lorsque les comportements de l'utilisateur reposent uniquement sur son intuition, « Décision délibéré » lorsqu'ils reposent uniquement sur ses raisonnements et « équilibré » lorsque les comportements résultent de la collaboration équitable de l'intuition et de la raison de l'utilisateur.

En réalité, les comportements d'un utilisateur sont toujours influencés à la fois par ses intuitions et par ses raisonnements. Néanmoins certaines situations impliquent une prédominance, plus ou moins importante, de l'un des aspects sur l'autre. La prévision ou l'identification de ces éventuelles prédominances pourrait permettre d'orienter la conception du système artificiel a priori, d'adapter le comportement du système en phase d'utilisation, ou d'analyser a posteriori ce qui s'est passé. C'est pour servir ces trois objectifs que nous dressons les archétypes de dispositions cognitives. Pour chaque disposition, nous énonçons des caractéristiques attendues à partir du modèle IPSEL et des caractéristiques des systèmes de traitement mobilisés.

4.2.2.1 Disposition équilibrée

Dans cette disposition en mode « équilibré » les comportements de l'utilisateur alternent entre automatismes intuitifs et actions délibérées à partir d'un raisonnement. Cet archétype illustre le comportement standard d'un opérateur humain qui dépend de son état émotionnel, de sa personnalité, de son savoir-faire et de son savoir. Bien qu'il puisse sembler être l'état cognitif le plus approprié à toute situation, cette disposition possède les défauts d'ouvrir la porte aux doutes, lorsque l'intuition et les raisonnements ne sont pas alignés, ou à l'influence négative d'un aspect cognitif sur l'autre. C'est par exemple le cas lorsque certains automatismes intuitifs biaisent le raisonnement, ou lorsque des raisonnements fallacieux supplantent des intuitions valables, ou encore lorsque les raisonnements interfèrent avec la mise en œuvre d'un contrôle moteur intuitif rapide et fluide. De façon simpliste, dans cette disposition les comportements de l'utilisateur sont polyvalents mais sous-optimaux.

4.2.2.2 Disposition Contrôle intuitif

La disposition cognitive « contrôle intuitif » illustre le cas d'un individu dont les comportements résultent uniquement des traitements de son système intuitif (S1). Dans cet état, le système délibératif (S2) n'est pas mobilisé. Les jugements, décisions et actions de l'utilisateur sont déterminés par son état émotionnel, sa personnalité et son savoir-faire. Ces comportements possèdent les caractéristiques inhérentes au système intuitif (S1), ils sont rapides, peuvent être opérés en parallèle, demandent relativement peu d'énergie et utilisent les connaissances implicites de l'individu. Cette disposition a l'avantage de permettre une prise de décision rapide, un contrôle moteur précis et fluide. L'inconvénient qui en découle est que l'utilisateur n'est pas disposé à raisonner conceptuellement. Par exemple, l'individu dans cette disposition aurait du mal à intégrer des connaissances explicites qui lui seraient fournies par le système comme dans le cas d'une explication ou d'arguments énoncés en langage naturel. De façon simpliste, l'utilisateur agit rapidement mais ne réfléchit pas.

4.2.2.3 Disposition Décision délibéré

Dans la disposition « Décision délibéré », c'est cette fois l'activité du système délibératif (S2) qui prédomine. Dans cet état, l'individu réfléchit efficacement et cela lui demande beaucoup d'énergie. Pour se faire il mobilise ses connaissances explicites, son savoir, et ses capacités de raisonnement logique. En conséquence, les fonctions du système intuitif (S1) sont amoindries. La perception de son environnement et sa conscientisation sont réduites,

le contrôle moteur est plus lent et moins précis. Dans cet état, l'individu est disposé à recevoir des explications et des arguments conceptuels mais ses réflexions peuvent éventuellement le couper du monde sensible et le rendre imperméable à certains signaux. De façon simpliste, l'utilisateur réfléchit beaucoup mais agit de manière inefficace et se coupe de ses sensations physiques.

4.2.2.4 Tableau récapitulatif

Nous synthétisons les caractéristiques des dispositions cognitives proposées dans le tableau ci-dessous.

Disposition cognitive	Connaissances mobilisées	Traitement cognitifs dominant	Avantages	Inconvénients
E - Equilibré	• Implicites (savoir-faire) • Explicites (savoir)	Aucun	Polyvalence des comportements	Sujet aux doutes. Raisonnement biaisés par l'intuition. Intuition amoindrie par les raisonnement.
C - Contrôle intuitif	• Implicites (savoir-faire)	Système Intuitif (S1) / Intuition	Perception et contrôle moteur précis, rapide et fluide. Décision rapide.	Jugements et décisions non délibérés conceptuellement. Sourd aux arguments et explications rationnels
D - Décision Délibéré	• Explicites (savoir)	Système Délibératif (S2) / Raisonnement	Jugements et décisions délibérés conceptuellement. Disposé à recevoir et à intégrer arguments et explications rationnels	Perception et contrôle moteur imprécis, lent et saccadés. Décision lente.

***Figure 4.2.3-2** : Dispositions cognitives d'un utilisateur et leurs caractéristiques.*

4.3 Analyse de l'interaction Homme-IA

Pour utiliser un système de la meilleure des manières, son utilisateur doit posséder certaines connaissances qui lui permettront d'évaluer la confiance à attribuer à son outil. Ces connaissances lui permettront de déterminer l'utilisation qui convient et de juger les productions dudit outil. En premier lieu, l'utilisateur doit connaître la fonction du système pour pouvoir évaluer si ce dernier l'aidera bel et bien à accomplir son objectif. L'utilisateur devrait également disposer de connaissances concernant le fonctionnement et les caractéristiques du système qu'il utilise pour évaluer si ce dernier pourrait effectivement

remplir sa fonction dans le contexte d'utilisation considéré. Ces connaissances permettent à l'utilisateur de prévoir les effets de l'utilisation, que ce soient des effets recherchés ou des effets de bord. Pendant l'utilisation, il est également important que l'utilisateur garde une certaine forme d'esprit critique. L'utilisateur doit pouvoir juger des productions du système qu'il utilise pour rester souverain de ses décisions et responsable de ses actes.

Ces connaissances de l'utilisateur, nécessaires pour la bonne utilisation de son outil, sont de différentes nature. Elles peuvent être acquises en amont, avant l'utilisation, ou pendant l'utilisation. Elles peuvent être implicites lorsqu'il s'agit de savoir-faire ou qu'elles ne sont pas conscientisées. Elles peuvent aussi être explicites, ce sont alors des savoirs à partir desquels l'utilisateur peut raisonner consciemment et délibérer son utilisation du système.

Pour analyser l'interaction entre un utilisateur et un système artificiel lui servant d'outil, nous proposons d'identifier différentes variables qui influencent l'acquisition et la constitution des connaissances de l'utilisateur. Nous distinguons deux phases; une première antérieure à l'utilisation et une seconde qui est la phase d'utilisation en elle-même. Pour cette seconde phase nous mobilisons les catégories de systèmes ainsi que les archétypes de dispositions cognitives que nous avons établis via la modélisation IPSEL. Dans une optique de décomposer pour mieux analyser, ces distinctions nous permettent de proposer des situations d'interactions Homme-Système caricaturales pour lesquelles nous discuterons l'état des variables qui influencent la bonne utilisation de l'outil.

4.3.1 Variables influençant la constitution des connaissances utilisateurs

Nous distinguons deux variables influençant la constitution des connaissances de l'utilisateur avant l'utilisation du système : l'instruction et l'entraînement.

L'instruction est un temps pendant lequel l'utilisateur reçoit un ensemble de connaissances générales portant sur la fonction du système, son fonctionnement, ses caractéristiques et des instructions concernant son utilisation. Durant ce processus, l'utilisateur reçoit essentiellement des connaissances explicites, c'est-à-dire un savoir, général.

A l'inverse, l'entraînement est un temps pendant lequel l'utilisateur construit un ensemble de connaissances implicites, un savoir-faire particulier. Durant l'entraînement l'utilisateur s'entraîne à utiliser un système particulier dans des conditions particulières.

Ces deux variables sont souvent liées mais elles n'en sont pas moins indépendantes. Il est possible de recevoir une instruction sans entraînement, tout comme il est possible de s'entraîner à utiliser un système sans avoir reçu d'instruction sur ce dernier.

Pour illustrer un peu mieux la différence entre instruction et entraînement nous prenons l'exemple d'un utilisateur et d'un système de navigation GPS. Durant l'instruction, l'utilisateur acquiert des connaissances explicites générales sur les systèmes de navigation GPS. Une entité tierce, que ce soit un formateur, un manuel d'instruction, des articles internet ou autres, lui apprennent que la fonction d'un système de navigation GPS est de proposer des itinéraires entre deux coordonnées GPS. L'utilisateur apprend que le système fonctionne en utilisant des cartes routières qui sont téléchargées dans le système qui adapte en continue l'itinéraire proposé en fonction de sa localisation GPS actuelle et de ses cartes. Une fois formé, l'utilisateur possède des connaissances générales sur les systèmes de navigation GPS, et ces connaissances lui permettent d'utiliser ce type de système de manière probablement plus appropriée que ne le ferait un utilisateur non formé.

Cependant toutes les situations d'utilisation de systèmes de navigation GPS, bien que ces systèmes aient la même fonction, le même fonctionnement et les mêmes caractéristiques générales, ne sont pas exactement identiques. Différents systèmes peuvent par exemple avoir des interfaces utilisateurs différentes. Leurs contextes d'utilisations ne sont pas non plus les mêmes. Utiliser un système de navigation lorsque l'on se déplace à pied ne demande pas d'effectuer exactement les mêmes comportements que lorsque l'on utilise ce même système en voiture. Ces savoir-faire particuliers et contextuels seront développés pendant l'entraînement.

Pour la variable « instruction », nous discriminons deux états qui qualifient la difficulté à former l'utilisateur sur le système considéré. Le premier état, « facile » sera attribué aux instructions qui concernent des systèmes dont la fonction et le fonctionnement sont facilement intelligibles pour un utilisateur néophyte. Le deuxième état, « difficile », est attribué aux instructions qui concernent des systèmes dont la fonction et le fonctionnement sont difficilement intelligibles pour un utilisateur néophyte.

Pour la variable « entraînement », nous discriminons deux états qui qualifient l'importance de la phase d'entraînement à l'utilisation du système considéré. Le premier état, « utile », est attribué aux entraînements qui sont utiles pour une meilleure utilisation du système considéré. Il s'agit d'un état par défaut car nous présupposons qu'il est toujours utile de

s'entraîner à utiliser un système. Le deuxième état, « critique », est attribué aux entraînements qui semblent essentiels pour une bonne utilisation du système considéré. Ce deuxième état correspond aux situations où l'instruction seule ne peut pas suffire, car le système ou les conditions d'utilisation sont si particulières que des connaissances générales ne permettent pas à elles seules de correctement utiliser le système.

Pour la seconde phase d'interaction, la phase d'utilisation en elle-même, nous distinguons deux variables influençant la constitution des connaissances qui permettront à l'utilisateur de juger et d'évaluer les productions du système : la contextualisation et l'explication.

La contextualisation correspond à une action menée par l'utilisateur qui contextualise les productions du système pour évaluer leur pertinence. Il s'agit pour l'humain de considérer la fonction, le fonctionnement et le domaine d'entraînement du système pour relativiser ses productions. Nous établissons deux états à cette variable : « utile » comme état par défaut, partant du principe qu'il est toujours utile pour l'utilisateur de relativiser les productions du système, mais que ce processus n'est pas essentiel compte tenu de la situation considérée ; et « critique » pour les situations où ce processus est essentiel pour la bonne utilisation du système considéré.

L'explication correspond à une communication du système qui fournit des éléments explicatifs à l'utilisateur. Intégrant ces éléments, l'utilisateur doit pouvoir comprendre les traitements qui ont amené le système à produire son résultat. Il s'agit pour le système de présenter un raisonnement causal explicitant les causes qui l'ont mené à générer les productions comme effets. Nous établissons deux états à cette variable : « utile » comme état par défaut, partant du principe qu'il est toujours utile pour l'utilisateur de recevoir des explications concernant les productions du système, mais que ce processus n'est pas essentiel compte tenu de la situation considérée ; et « nécessaire » pour les situations où ce processus est essentiel pour l'interaction considérée.

Nous synthétisons l'ensemble des variables intervenant dans notre analyse de l'interaction entre l'Homme et un système artificiel :

- l'instruction de l'utilisateur, en phase de pré-utilisation, pouvant être facile ou difficile.
- l''entraînement de l'utilisateur, en phase de pré-utilisation, pouvant être utile ou critique.

- la contextualisation des productions du système par l'utilisateur, pendant l'utilisation, pouvant être utile ou nécessaire.
- l'explication des productions du système par le système lui-même, pendant l'utilisation, pouvant être utile ou nécessaire.

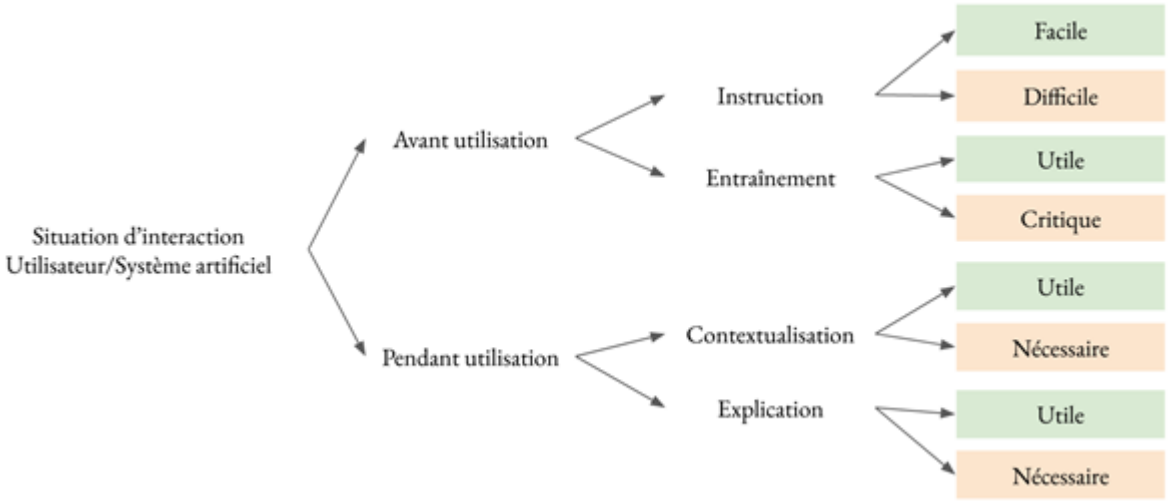

***Figure 4.3.1** : Variables qualifiant les prérequis à une bonne utilisation d'un système artificiel.*

4.3.2 Qualification des situations d'interaction Utilisateur - Système artificiel

En utilisant ces variables, nous proposons maintenant de qualifier les situations d'interactions entre un utilisateur et un système artificiel. Pour cela nous supposons l'état de ces différentes variables en fonction de la catégorie du système et de la disposition cognitive de son utilisateur. Ce faisant, nous établissons une liste d'éléments contraignant l'interaction entre un utilisateur humain et son outil informatique.

La situation la moins contraignante est celle où il est facile d'instruire l'utilisateur, où la phase d'entraînement en condition est utile mais pas critique, où la contextualisation des productions du système par l'utilisateur est utile mais pas nécessaire et où l'explication des productions du système par le système lui-même est utile mais pas nécessaire.

A l'opposé la situation la plus coercitive est celle où il est difficile de former l'utilisateur, où la phase d'entraînement en condition est critique, et où la contextualisation des productions par l'utilisateur et l'explication des productions par le système sont absolument nécessaires.

Entre ces deux situations extrêmes se trouvent une multitude de cas particuliers que nous proposons de regrouper dans des archétypes de situations caricaturales dépendantes de la catégorie du système concerné et la disposition cognitive de l'utilisateur qui l'utilise.

Afin de décrire notre analyse de manière didactique nous proposons de remplir pas à pas le tableau ci-dessous:

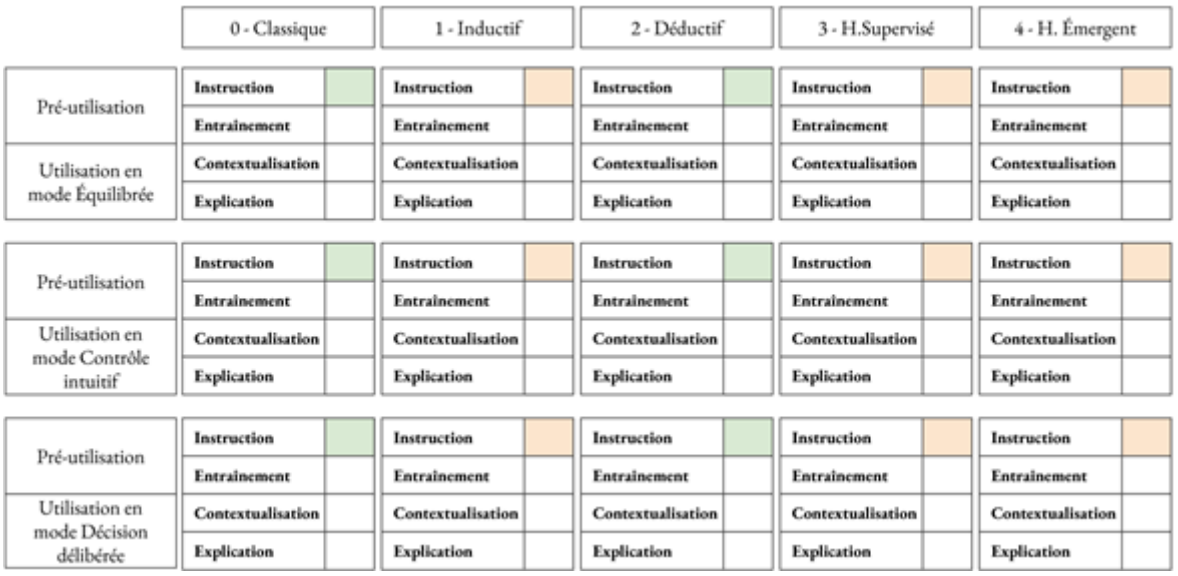

***Figure 4.3.2** : Situations d'interactions Homme-Système artificiel en fonction de la disponibilité cognitive de l'utilisateur et de la catégorie du système.*

Variable « instruction » : L'instruction ayant lieu durant la phase de pré-utilisation, sa difficulté est indépendante de la disposition cognitive de l'utilisateur pendant l'utilisation et ne dépend donc que de la catégorie du système. Nous présumons « facile » l'instruction concernant les systèmes non-intelligents de catégorie *Classique,* car leurs fonctions, fonctionnements et productions générées sont prévus lors de la conception du système et ne sont pas modulés par un quelconque apprentissage. Nous présumons également « facile » l'instruction concernant les systèmes intelligents de la catégorie *Déductif,* car la modulation de leurs comportements initialement prévus sont dirigés par des règles qui ont été données au système par l'Homme. Ces règles peuvent donc être explicitées lors de l'instruction. Nous présumons « difficile » l'instruction concernant les systèmes intelligents dont tout ou partie des comportements peuvent être modulés par un apprentissage statistique. Pour ces systèmes, qui sont donc ceux de catégorie 1, 3 et 4, leurs fonctions, fonctionnements et productions ont le potentiel d'être modifiés en fonction des données que le système particulier traitera durant son apprentissage.

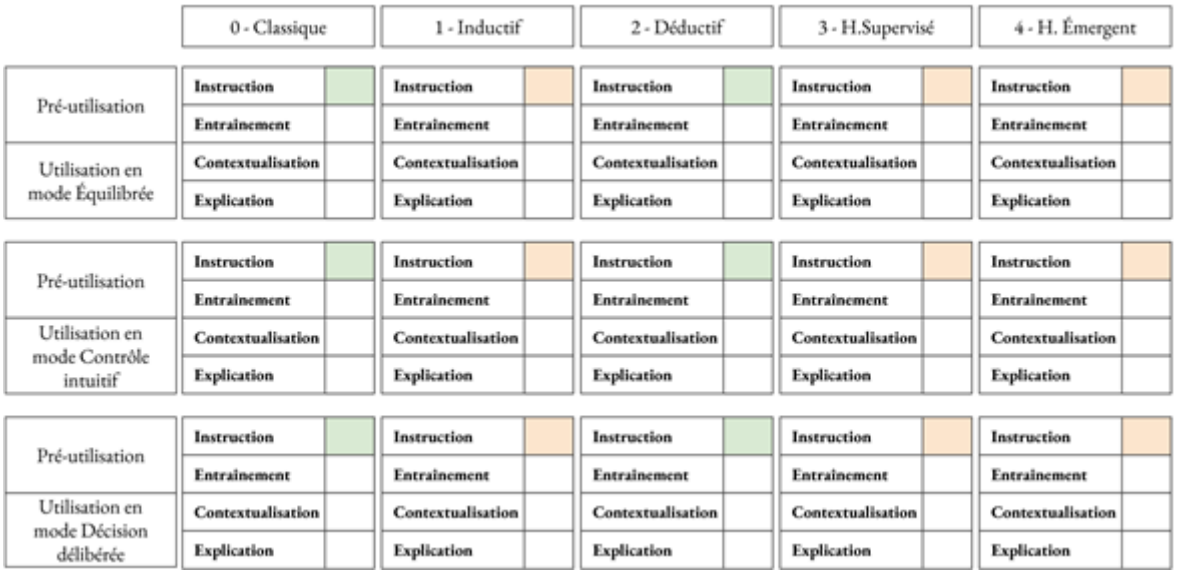

Figure 4.3.2-2 *: Présomptions de l'état de la variable « instruction ».*

Variable « entraînement » : Une phase d'entraînement est toujours utile pour servir la bonne utilisation d'un système. Cependant cette phase peut être considérée comme nécessaire pour certaines situations particulières, en fonction de la catégorie du système concerné et de la disposition cognitive de l'utilisateur lors de la phase d'utilisation.

Nous présumons « critique » la phase d'entraînement correspondant aux situations où l'utilisateur utilise le système dans la disposition cognitive « Contrôle intuitif », puisque c'est essentiellement son savoir-faire qui guidera son utilisation. L'entraînement semble également nécessaire quand les comportements du système dépendent de son apprentissage particulier et ne peuvent donc pas être totalement appréhendés lors de l'instruction, comme cela est le cas pour les systèmes de catégorie « Inductif », « Hybride Supervisé » et « Hybride Emergent ». Cela est d'autant plus nécessaire que cette même caractéristique nous avait déjà amené à considérer difficile l'instruction de l'utilisateur pour ces catégories, nous définissons donc comme « critique » l'entraînement à l'utilisation des systèmes des catégories 1, 3 et 4.

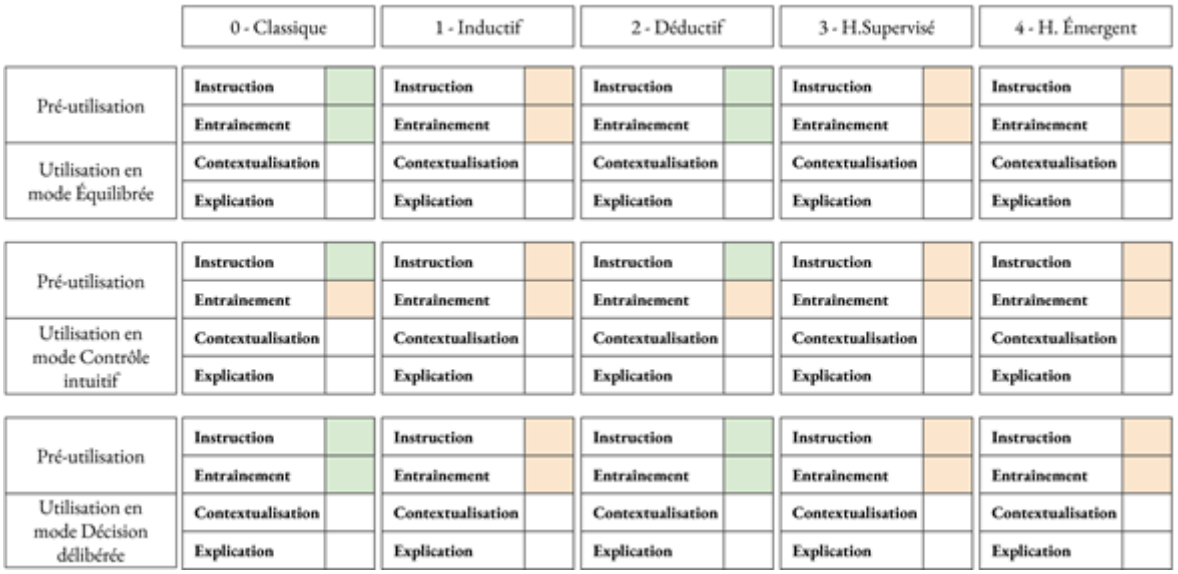

Figure 4.3.2-3 *: Présomptions de l'état de la variable « entraînement ».*

Variable « contextualisation » : Que l'utilisateur contextualise les productions du système qu'il utilise semble toujours utile. Cependant ce processus peut devenir nécessaire pour les situations où les productions du système peuvent être biaisées par son domaine d'apprentissage, c'est-à-dire lorsque le système effectue un apprentissage statistique sans être confronté à une logique humaine qui pourrait identifier des biais de représentativité et les corriger. Ces situations sont celles où le système ne possède pas de capacité de raisonnement logique (catégorie 1) ou celle où la logique employée par le système n'est pas une logique humaine mais une logique émergente, potentiellement non alignée avec celle de l'utilisateur (catégorie 4). Il faut néanmoins retrancher (remettre dans l'état par défaut) les situations où l'utilisateur n'est pas disposé à effectuer ce processus de contextualisation, c'est-à-dire lorsqu'il est dans la disposition « Contrôle intuitif ».

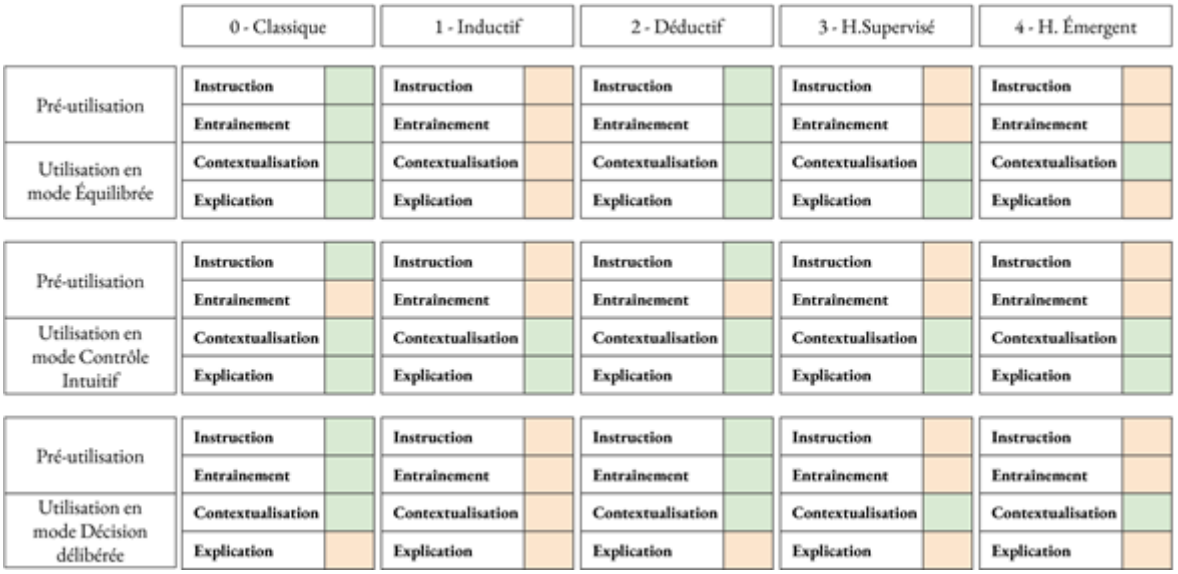

Figure 4.3.2-4 *: Présomptions de l'état de la variable « contextualisation ».*

Variable « explication » : Que le système explique ses productions semble toujours utile. Cependant ce processus peut devenir nécessaire pour les situations impliquant un utilisateur dans la disposition « Décision délibérée », car ses décisions reposent essentiellement sur des raisonnements délibérés qui se nourrissent d'arguments explicites. Ces justifications des productions par le système sous forme d'explication semblent également nécessaires lorsque le système a la capacité de produire des productions qui n'ont pas été prévues lors de la conception (catégorie intelligente 1, 2, 3, 4) mais qu'aucun raisonnement logique de la part du système ne les a validées (catégorie 1) ou que le raisonnement n'emploie pas une logique humaine (catégorie 4). Il faut également retrancher les situations où l'utilisateur n'est pas en capacité de recevoir ces explications comme c'est le cas lorsqu'il est dans la disposition « Contrôle intuitif ».

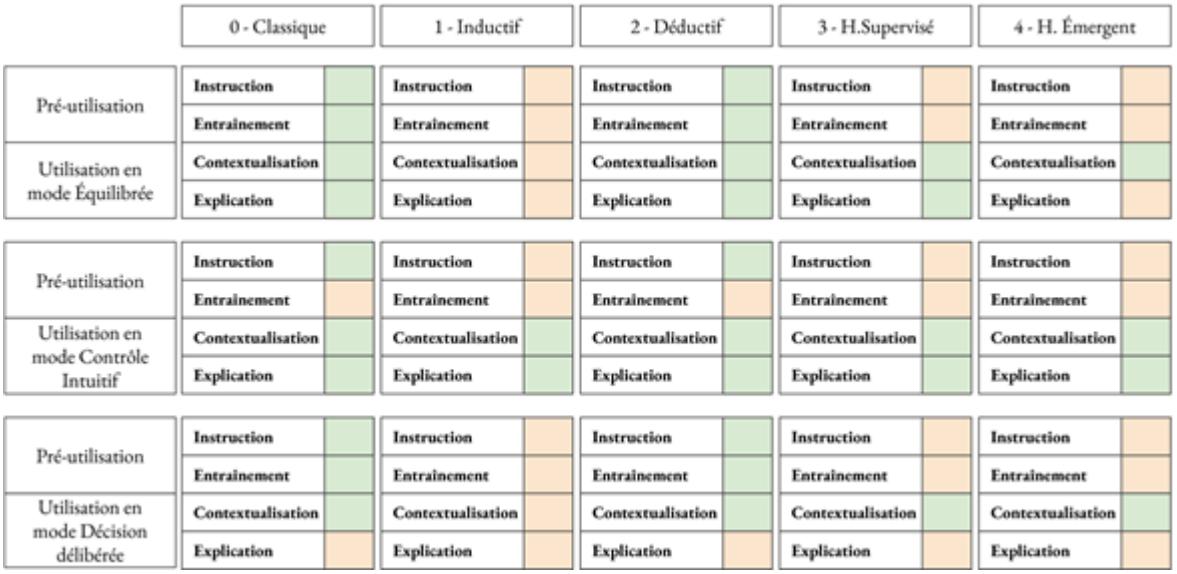

Figure 4.3.2-5 *: Éléments coercitifs situationnels à une bonne interaction Homme-IA.*

4.3.3 Interprétation

Toute situation d'interaction entre un utilisateur et un système artificiel peut être améliorée en permettant à l'utilisateur de disposer des connaissances nécessaires à la bonne utilisation de son outil. Nous avons établi quatre variables influençant l'acquisition de ces connaissances : l'instruction, l'entraînement, la contextualisation des productions et l'explication des productions.

Les situations d'interaction entre un utilisateur et un système de catégorie 0-Classique sont celles qui présentent le moins d'éléments coercitifs. Cela est dû au fait que ces systèmes n'apprennent pas et que leurs comportements sont donc plus facilement prévisibles. Cependant dans le cas d'une utilisation dans la disposition « Contrôle intuitif » il serait nécessaire que l'utilisateur effectue un entraînement en condition. Dans le cas d'une utilisation dans la disposition « Décision délibérée » il serait nécessaire que le système puisse expliquer ses productions, une prérogative relativement accessible puisque le système utilise une logique sans apprentissage (donc immuable) définie par l'Homme lors de la conception.

Néanmoins les systèmes de catégorie 0-Classique, proposés comme analogues au système direct (S0) de l'architecture IPSEL, ont des capacités qui se limitent à celles prévues lors de leur conception. Dépourvu de mécanisme d'apprentissage, un tel système s'adapte mal aux situations et aux problématiques complexes. Les systèmes de catégories 1-Inductif et

2-Déductif, analogues respectivement aux systèmes intuitif (S1) et délibératif (S2) de l'architecture IPSEL, sont plus performants et s'adaptent mieux aux situations complexes du fait de leur capacité d'apprentissage.

Les situations d'interaction entre un utilisateur et un système de catégorie 1-Inductif sont celles qui présentent le plus d'éléments coercitifs. Cela est dû au fait qu'il s'agit là de systèmes intelligents d'une part, et que leur intelligence utilise des connaissances implicites acquises par généralisation de leur expérience d'autre part. Pour ces systèmes la phase d'entraînement est primordiale puisque, du fait de cet apprentissage inductif, chaque système possède les connaissances subjectives inhérentes à son domaine d'apprentissage particulier. Ne disposant pas de capacité de raisonnement logique, ces systèmes peinent à expliquer leurs productions autrement que statistiquement, ils sont donc difficiles à utiliser pour un utilisateur dans la disposition « Décision délibérée ». Ces systèmes semblent plus adaptés à une utilisation dans la disposition « Contrôle intuitif », certainement pour des tâches relatives aux perceptions et aux contrôles.

Les situations d'interaction entre un utilisateur et un système de catégorie 2-Déductif, présentent peu d'éléments coercitifs. Cela est dû au fait que leur nature est assez similaire à celles des systèmes de catégorie 0-Classique, excepté qu'ils ont la capacité de déployer un raisonnement logique pour découvrir de nouvelles connaissances explicites à partir d'inférence logique. Il est donc relativement facile d'interagir correctement avec ce type de systèmes mais leurs performances sont limitées par la capacité de l'Homme à modéliser explicitement leurs domaines d'opération, ce qui peut s'avérer très difficile pour les environnements complexes. On pourrait donc dire qu'il est toujours plus intéressant d'utiliser un système déductif qu'un système classique, mais que la conception d'un système déductif est plus difficile.

Les systèmes de catégorie 1 et 2 sont plus performants que les systèmes de catégorie 0 du fait de leur capacité à apprendre. Néanmoins leur utilisation présente des inconvénients notables tels que leur subjectivité dans le cas des systèmes inductifs, ou leur difficulté de conception dans le cas des systèmes déductifs. Pour mitiger ces inconvénients il est de plus en plus courant de faire coopérer ces deux types d'apprentissage dans des systèmes qualifiés d'Hybrides.

Les situations d'interaction entre un utilisateur et un système de catégorie 3-Hybride_supervisé présentent moins d'éléments coercitifs que les situations impliquant

des systèmes de catégorie 2. Cela est dû à l'apport de l'hybridation de l'apprentissage inductif avec des capacités de raisonnement logique. Néanmoins, si cette hybridation permet de mitiger l'inconvénient de la subjectivité des connaissances du système, elle ne lève pas l'inconvénient de la difficulté de modéliser, dans une logique humaine, l'environnement opérationnel tel que cela est le cas pour la conception des systèmes déductifs.

L'hybridation des systèmes de catégorie 4-Hybride_émergent mitige à la fois l'inconvénient de la subjectivité des connaissances, grâce à l'encapsulation de ces connaissances dans un raisonnement logique, et elle mitige également la difficulté de conception puisque les logiques qui modélisent l'environnement opérationnel n'ont pas à être définies par l'Homme mais sont construites par le système lui-même. Cependant, cette auto-construction de la logique entraîne d'autres inconvénients particuliers engendrant de nouveaux éléments coercitifs aux situations d'interaction les impliquant. Il s'agit principalement de permettre à l'utilisateur et aux systèmes de s'assurer que leurs connaissances explicites et leurs logiques d'inférence sont alignées et cohérentes les unes avec les autres.

4.4 Discussions à propos de l'IA de Confiance

Le développement de l'intelligence artificielle, et de son utilisation dans de nombreux secteurs d'activités, amène tout un chacun à s'interroger sur la confiance à accorder dans ces nouvelles technologies.

D'après un rapport effectué dans le cadre du « World Economic Forum » [Tazrout 2022], 60% des personnes interrogées sont d'accord pour dire que les produits et services utilisant l'I vont rendre leur vie plus facile, mais 40% admettent également que cette technologie les rend nerveux et les inquiète. 50% des personnes interrogées disent ne pas faire autant confiance aux entreprises qui utilisent l'IA qu'aux entreprises qui ne l'utilisent pas.

Bien que les technologies de l'IA aient permis de nombreux accomplissements, leurs usages comportent des risques mis en lumière par différents chercheurs et organisations internationales [Bostrom 2017] [Brundage et al. 2018]. Pour tenter de minimiser ces risques et protéger les individus et les sociétés, plusieurs entités gouvernementales ont publié ces dernières années des rapports, des feuilles de route et des cadres juridiques dans le but d'encadrer et de réguler le développement et l'utilisation des systèmes dotés d'IA.

[Villani et al. 2018] [Macron 2018] [anonyme 2021]. L'objectif de tous est de développer une « IA de confiance », ou « Trustworthy AI » en anglais.

Pour le chercheur Freddy Lecue, une IA de confiance est une IA explicable, responsable et qui protège la vie privée [Lecue 2019]. Nous dirons que les attributs « responsable » et « qui protègent la vie privée » sont des considérations éthiques. Une IA de confiance serait alors un système utilisant des techniques provenant du domaine de recherche de l'intelligence artificielle, et ce système aurait en plus les caractéristiques d'être explicable et éthique.

4.4.1 IA Explicable

Les acteurs du domaine de l'IA ont le souci commun de rendre les algorithmes plus transparents en dotant ces algorithmes de capacités d'explication de leurs comportements et de leurs résultats. L'hypothèse sous-jacente est de dire que si les utilisateurs comprennent mieux les systèmes, leurs fonctionnements et leurs résultats, ils auront plus confiance envers ces derniers, ou du moins ils pourront leur faire confiance. Il existe plusieurs façons de produire des explications, il peut s'agir d'une justification du comportement d'un agent artificiel, du débogage d'un modèle statistique, de l'argumentation d'une décision médicale ou de l'éclaircissement de prédictions. Les travaux œuvrant sur ce chemin de l'explicabilité sont usuellement attachés au domaine nommé X.A.I. pour « explainable artificial intelligence ».

De nombreux chercheurs comme Freddy Lecue travaillent sur la problématique de l'explicabilité. Certains reprochent à cette communauté de recherche, encore jeune, de majoritairement utiliser leurs propres intuitions sur ce que doit être une bonne explication et ne mobilisent pas suffisamment les connaissances provenant des sciences sociales [Miller et al. 2017]. L'appréciation des explications est en effet un sujet complexe puisque cette appréciation est subjective et dépend des connaissances de l'individu qui les reçoit [Ehsan et al 2021]. Qui plus est, les systèmes informatiques manipulent des chiffres et un modèle numérique du monde. Or il a été montré que les individus humains sont moins enclins à accepter des explications concernant des chiffres plutôt que des explications exprimées en langage naturel [Miller 2019] [Kahneman 2011] [Ehsan et al 2021] et de nombreux aspects de la réalité ne peuvent pas être modélisés par des chiffres. Le problème réside alors dans le fait que, malgré les récentes performances de supers modèles de langage naturel comme GPT-3, la communauté s'accorde à dire que les machines ne comprennent pas le langage

naturel. Dans ce sens, Ciravegna et al. [2021] ont travaillé sur des méthodes permettant de fournir des explications avec des formules de logique du premier ordre, un pas vers un langage plus interprétable que celui des chiffres et moins abstrait que le langage naturel. D'autres directions sont explorées. Des méthodes algorithmiques, comme SHAP ou LIME, ont émergé et permettent d'avancer vers une meilleure explicabilité des modèles de machine learning [Ribeiro et al. 2016] [Nori et al 2019].

Les expériences empiriques [Wiegmann 2002] et le sens commun nous montrent qu'une explication n'est pas le seul critère qui intervient dans la construction de la confiance par un individu. Pour illustrer ce propos nous prendrons l'exemple du projet de conception d'un système d'aide à la décision à destination d'experts en cardiologie d'un hôpital américain. Le professeur Zimmermann, responsable de ce projet, rapporte que les cardiologues ont développé une grande confiance personnelle à travers leurs propres expériences et les résultats qu'ils ont obtenus au fil de leur carrière. Cette confiance leur est nécessaire pour agir avec détermination, lever les doutes et « avoir la main sûre ». Pour ce type de profil, « experts » dans leur domaine, Zimmerman observe une méfiance naturelle envers les préconisations fournies par la machine, quand bien même elle ajouterait une explication logique et cohérente [Zimmerman 2018]. Dans ce sens, certains chercheurs remettent en question l'intérêt des explications a posteriori qu'ils jugent insatisfaisantes voire dangereuses. Elles pourraient laisser croire que l'on peut faire confiance au système alors que l'explication de son comportement, bien que convaincante, est potentiellement fausse [Rudin 2019]. Si l'explication permet de donner confiance dans le système, qu'est ce qui donne confiance dans l'explication ? Cynthia Rudin [2019] par exemple prône une meilleure interprétabilité des processus en modifiant la construction des programmes plutôt que le développement de programmes complémentaires dont l'objectif est d'expliquer les résultats des premiers [Chen et al. 2020].

L'explicabilité n'est pas sans défis. Plus encore, l'explicabilité ne peut être considérée comme la seule variable affectant la confiance de l'utilisateur. Il est nécessaire de prendre en compte l'individu dans son ensemble, sa personnalité, son expérience, ses émotions et ses instincts lors de l'analyse de l'interaction Homme-Machine. Ces notions de facteurs humains, évoqués dans la problématique de l'X.A.I. par la condition « explications compréhensibles par l'Homme », appellent à une réflexion bilatérale. Des questions subsidiaires comme « Qu'est ce qui est compréhensible pour l'Homme » ou « comment s'assurer de la bonne compréhension de l'Homme » doivent légitimement entraîner une étude qui ne peut pas se

borner aux seuls aspects technico-algorithmiques des technologies mais doit également inclure les mécanismes cognitifs de l'Homme.

4.4.2 IA Éthique

Pour illustrer les problématiques éthiques liées à l'IA nous rapportons les travaux de Timnit Gebru. Cette chercheuse travaillant sur l'éthique des systèmes intelligents s'est fait connaître notamment par son éviction de Google, son employeur d'alors. Chez Google, son équipe a rédigé une publication dans laquelle elle dénonce les risques sociétaux que font courir la conception et l'utilisation de modèles d'IA servant au traitement du langage naturel [Hao 2020]. Pour l'équipe de chercheurs, ces risques sont au nombre de quatre. (i) Il y a tout d'abord le fait que la conception et l'entraînement de tels modèles consomment beaucoup d'énergie électrique débouchant sur une empreinte carbone considérable et d'importants coûts financiers [Strubell 2019]. En conséquence, le groupe de recherche suggère que l'utilisation de ce type de modèle va principalement bénéficier aux organisations « riches », alors que les conséquences du changement climatique frappent plus durement les communautés pauvres et marginalisées, ce qui pose un problème d'inégalité. (ii) Deuxièmement, ces modèles sont entraînés sur d'immenses volumes de données, ce qui signifie que leurs concepteurs récupèrent et utilisent toutes sortes de textes provenant d'Internet. Il est alors difficile, voire impossible, de contrôler toutes ces données en amont. Le risque est donc que le modèle intègre des formes d'expressions racistes, sexistes ou abusives. De plus, le modèle rencontrera des difficultés à prendre en compte le langage des communautés ayant le moins accès à Internet, créant ainsi des discriminations. En conséquence le langage artificiel ainsi généré sera homogène et représentera les pratiques des pays et communautés les plus riches. (iii) Le troisième risque relevé par l'équipe Timnit Gebru est nommé « Misdirected research effort ». Bien que la plupart des chercheurs dans le domaine du NLP affirment que les modèles de langage automatique ne saisissent pas le sens des phrases mais sont excellents dans la manipulation de leur forme [Heaven 2020], les sociétés informatiques génèrent beaucoup d'argent en les utilisant. Les industriels continuent à investir sur les techniques qui imitent la forme du discours plutôt que d'investir dans la recherche de modèles qui seraient capables d'en comprendre le fond. En somme, la communauté développe prioritairement des systèmes rentables mais ayant le potentiel de tromper les humains en manipulant un langage incompréhensible pour les machines. (iv) Dernier point, ces modèles de langage sont devenus suffisamment bons pour créer des phrases, paragraphes et textes complets qui imitent parfaitement la forme

d'expression humaine. Il est donc désormais facile pour ceux qui les contrôlent de générer de grandes quantités d'information pour manipuler les peuples et les désinformer. D'après Mme Gebru, cela expose les démocraties et l'organisation sociétale à des manipulations ou des actions d'ingérence.

La problématique de l'éthique est globale et semble être une affaire de société. Il est possible pour les gouvernements et institutions juridiques de s'accorder sur des standards et des bonnes pratiques dont l'objectif serait d'assurer que certains risques sont mitigés en amont lors de la conception des systèmes. Dans ce sens, la commission européenne a publié un règlement pour harmoniser les pratiques législatives dans chaque Etat, concernant la mise sur le marché et l'exploitation des systèmes d'IA [anonyme 2021-2].

4.4.3 « IA de confiance »

Une IA de confiance est pour le moment considérée comme une IA éthique et capable de donner une explication à ses comportements. La notion d'explicabilité est pertinente mais complexe et ne saurait trouver de solution universelle puisque les situations sont diverses et les connaissances des individus subjectives. Il en est de même pour l'aspect éthique des systèmes. L'éthique étant une construction socio-culturelle, il semble difficile de pouvoir concevoir des systèmes universellement éthiques. Pour avancer vers une subjectivité commune, il sera nécessaire d'établir des normes et des lois encadrant le développement de l'IA. L'utilisation d'IA de confiance requiert d'identifier clairement les standards selon lesquels un système peut être qualifié « de confiance », de créer des institutions capables de juger du bon respect de ces standards et de développer des méthodes qui intègrent ces standards de manière durable.

De notre point de vue, cette analyse de la confiance est faite en considérant principalement le système artificiel. C'est à la machine que l'on cherche à apposer des caractéristiques, celles d'être explicable et éthique, en délaissant peut être les facteurs humains. Dans cette vision il nous semble que l'hypothèse est de définir que si le système explique ses comportements et a été conçu en suivant un ensemble de règles permettant de le qualifier d'éthique, alors l'utilisateur peut, et aura confiance en la machine. Or les observations empiriques montrent que les explications ne suffisent pas toujours [Zimmerman 2018], et que l'éthique est une construction socio-culturelle. Il nous semble donc intéressant d'analyser l'interaction entre individus et systèmes artificiels par le cadre des sciences cognitives et des facteurs humains. Pour cela nous mobiliserons la modélisation que nous avons proposée et qui nous

permettra de dresser un tableau des relations entre la confiance de l'individu et ses intéractions avec la machine.

La confiance est un état physio-psychologique de l'utilisateur qui détermine sa propension à utiliser l'outil. Cet état du corps et de l'esprit est engendré par l'évaluation et le jugement que se fait l'utilisateur de la situation et de l'outil, au regard de ses objectifs. Inconsciemment et consciemment, l'utilisateur évalue et juge la situation, les entités qu'elle fait intervenir et les risques qui en découlent potentiellement. Le système artificiel et son état sont évalués. Ces jugements portent sur les caractéristiques du système, sa conception, ses capacités, ses limites, son fonctionnement et les performances que l'individu peut en espérer. Ils sont également nourris des informations implicites et explicites qui sont interprétées à partir du comportement de l'outil lors de l'utilisation.

L'ensemble de ces activités cognitives va permettre à l'utilisateur de calibrer son état de confiance, et ainsi guider sa décision d'utiliser ou non, le système à sa disposition.

Nous montrerons ici que l'explicabilité et l'éthique en tant qu'attributs que l'on peut apposer au système ne sont pas les seules variables qui entrent dans la composition des jugements et des traitements cognitifs déterminant la confiance de l'individu. Une approche globale permettra d'identifier les facteurs à considérer pour s'assurer de la bonne interaction entre utilisateurs et systèmes artificiels.

Pour raffiner notre analyse, il nous faut définir deux aspects de la confiance envers un système artificiel. Il y a en premier lieu la confiance attribuable à la conception du système : ce qui revient pour l'individu à faire confiance aux constructeurs, aux organismes de normalisation et de certification. Cette confiance pourrait être nommée confiance systémique. Il y a aussi, indépendamment de cette confiance systémique, une confiance situationnelle, qui se calibre lors de l'utilisation, au fil des intéractions particulières entre l'individu et son outil.

4.4.4 Confiance systémique et confiance situationnelle

Un premier volet de la confiance concerne la confiance dans la construction, l'entraînement et les tests qui ont été menés en respectant les normes et bonnes pratiques en vigueur dans la/les sociétés où les systèmes sont utilisés. Ces standards, encadrant la conception des systèmes, ont été établis par des institutions gouvernementales, en collaboration avec les

institutions académiques, industrielles et législatives. Ils assurent que la conception des systèmes est alignée avec les valeurs éthiques des sociétés dans lesquelles ils seront utilisés et mitigent les risques associés à leur utilisation.

Un second volet de la confiance concerne la confiance de l'individu lors de l'utilisation d'un système particulier, dont l'usage et les comportements permettent à l'utilisateur de correctement calibrer la confiance qu'il peut attribuer au système pour l'utilisation situationnelle qui est la sienne. Cette confiance active guide l'utilisation du système et permet à l'utilisateur de garder la souveraineté de ses décisions et la responsabilité de leurs conséquences par rapport à la situation d'utilisation qui est la sienne.

Cette distinction nous semble importante car confiance systémique et confiance situationnelle sont complémentaires. Un système dont la conception est certifiée « de confiance » pourrait, lors de l'utilisation, ne pas permettre à l'utilisateur de correctement calibrer la confiance à lui attribuer pour la situation spécifique qui est la sienne, entraînant une utilisation involontairement préjudiciable pour l'utilisateur, un tiers physique ou moral, ou la société. Inversement, l'utilisation d'un système qui donne confiance à son utilisateur pourrait néanmoins entraîner des conséquences préjudiciables si sa conception le permet. Précisons ici que l'objectif n'est pas de grandir artificiellement la confiance situationnelle de l'utilisateur. L'excès de confiance peut entraîner de mauvaises utilisations, tout comme le défaut de confiance. Pour prendre de bonnes décisions, notamment sur l'emploi ou non de son outil, l'utilisateur doit avoir une confiance situationnelle appropriée et ne peut se contenter d'une confiance systémique.

De manière générale, pour que l'emploi des systèmes d'IA soit bénéfique, il faut que ces systèmes soient bien conçus (Confiance dans la conception, confiance systémique) et bien utilisés (Confiance appropriée lors de l'utilisation, confiance situationnelle).

4.4.5 Conception de confiance

La bonne conception de systèmes d'IA nécessite la mise en place de normes et de bonnes pratiques assurant que la construction, l'entraînement et les tests du système suivent certains standards certifiant que ce dernier peut être qualifié « de confiance ».

Ces standards doivent être établis par des institutions disposant de l'autorité et des connaissances nécessaires. Il peut s'agir d'organismes de normalisation, d'instances

gouvernementales ou de commissions ad hoc. Ils peuvent comporter des prérequis et obligations concernant l'acquisition et le traitement des données d'entraînement. Comme par exemple leur traçabilité, les méthodes d'anonymisation appliquées, le lieu et la période de stockage etc. Les standards devraient aussi établir des objectifs quant aux processus d'entraînements, par exemple des scores de précisions à atteindre ou des évaluations de la représentativité de l'entraînement par rapport au domaine d'application. Enfin, les prérequis doivent aussi concerner les algorithmes et leurs propriétés comme leur transparence, explicabilité et interprétabilité.

Néanmoins, ces standards ne pourront être considérés autrement que comme des constructions humaines comportant une part d'abstraction et de subjectivité. Par exemple, l'éthique dans les pays occidentaux n'est pas la même que celle dans les pays orientaux, il se peut que des pays de cultures différentes établissent des standards différents. L'exemple le plus parlant actuellement est celui de la protection des données privées qui n'est pas abordé de la même manière en Europe et en Chine. Qui plus est, les techniques de l'intelligence artificielle évoluant sans cesse, il sera difficile d'établir des mesures et méthodes générales, qui resteront applicables dans 5, 10 ou 15 ans. Notons également que ces normes ne seront pas nécessairement applicables uniformément à tous les secteurs d'activités. Par exemple, le règlement du parlement européen [anonyme 2018-2] prévoit un certain nombre de règles à respecter pour la mise sur le marché et l'exploitation de systèmes dotés d'intelligence artificielle, mais leurs applications dépendent en partie du secteur d'activité concerné. Par exemple, ce règlement précise que « Les systèmes d'IA exclusivement développés ou utilisés à des fins militaires devraient être exclus du champ d'application du présent règlement lorsque cette utilisation relève de la compétence exclusive de la politique étrangère et de sécurité ».

Ces ensembles de normes et de bonnes pratiques certifiant les systèmes d'intelligence artificielle en tant que système de confiance sont une bonne avancée vers une meilleure intégration des technologies de l'IA dans nos activités humaines. Cependant ils seront difficiles à établir de manière générale et stable. Normes et bonnes pratiques dépendent des sociétés et des cultures qui les érigent ainsi que des techniques algorithmiques qui évoluent encore beaucoup dans ce domaine. Par ailleurs, elles ne parviendront pas à éviter des utilisations volontairement préjudiciables. La certification ne concerne que la conception et ne pourra jamais empêcher une utilisation détournée des systèmes à des fins malveillantes. Il se pourrait également que les systèmes, bien que certifiés de confiance, soient

involontairement mal-utilisés, entraînant des préjudices tout aussi importants aux individus ou aux sociétés. La bonne conception n'assure pas la bonne utilisation.

4.4.6 Utilisation en confiance

Les systèmes d'IA ont la particularité d'apprendre, c'est-à-dire de changer leurs comportements. Les mécanismes à l'œuvre sont complexes et souvent inintelligibles pour l'utilisateur. La bonne conception de système d'IA doit être accompagnée de leur bonne utilisation. Cette utilisation appropriée passe par une confiance correcte de l'utilisateur envers son outil, pour la situation qui est la sienne. L'outil peut-il m'aider à accomplir mon objectif ? Quelles sont ses capacités et ses limites ? Quelles sont les conséquences de son utilisation ? Ces questions doivent être adressées par l'utilisateur s'il veut garder la souveraineté de ses décisions et devoir assumer la responsabilité de ses actes. Pour cela, il est nécessaire que l'outil lui-même, spécialement dans le cas de systèmes d'IA, communique un ensemble d'informations qui permettront à l'utilisateur d'agir en connaissance de cause. Un tel outil pourrait être qualifié comme « donnant confiance ».

Durant notre travail nous avons proposé une architecture cognitive ayant l'objectif de donner du sens dans les relations entre jugements, connaissances, décisions, actions et émotions. Nous la mobilisons maintenant pour analyser l'interaction entre un utilisateur et son outil d'intelligence artificielle via ce spectre de la confiance.

L'état de confiance est la résultante d'une évaluation et d'un jugement. Comme nous l'avons postulé à travers notre modélisation, ces jugements peuvent être de deux natures : (1) Implicites quand ils sont une production du système intuitif et mobilisent des connaissances implicites ; (2) Explicites quand ils sont une production du système déductif et mobilisent des connaissances explicites. Pour déterminer l'état de confiance d'un individu, il serait alors nécessaire de connaître son expérience, ses perceptions et ses connaissances. Seulement, comme nous l'avons également postulé, les connaissances d'un individu sont subjectives et individuelles.

Il semble donc impossible de prédire la confiance qu'aura un individu dans une situation donnée. Car nous ne saurons jamais précisément ce qu'il perçoit, ce qu'il sait, quelle est sa personnalité, dans quel état émotionnel et psychologique il se trouve, etc.

Pour autant, le modèle IPSEL, étant donné les liens qu'il décrit entre les diverses entités du système cognitif, nous permet d'identifier les variables qualitatives qui entrent en jeu dans le mécanisme d'évaluation de la confiance. En appliquant le principe de « dualité-cognitive » que nous avons proposé, nous établirons que la confiance globale d'un individu peut être considérée comme résultant de l'évaluation de deux confiances sous-jacentes : (1') la confiance intuitive, résultat des traitements du système intuitif, et (2') la confiance délibérative, résultat des traitements du système délibératif.

Si nous ne pouvons pas prédire les résultats de ces traitements, nous pouvons supposer les moyens de les rendre les plus pertinents possibles. Le modèle stipule que les traitements du système intuitif sont déterminés par les perceptions, les émotions et les connaissances implicites alors que ceux du système délibératif sont déterminés par la conceptualisation de l'état du monde et les connaissances explicites. Pour rappel les connaissances implicites sont un ensemble de connaissances apprises lentement au cours de l'expérience et contiennent le savoir-faire. Les connaissances explicites quant à elles dépeignent une représentation du monde courant et de ses dynamiques sous la forme de concepts abstraits partageant des caractéristiques et des fonctions. Elles contiennent le « savoir » et peuvent être acquises rapidement via un média tel que le langage.

En dressant un tableau général de la situation d'interaction entre un utilisateur et un outil d'IA, nous dirions qu'une utilisation appropriée de l'outil est hautement plus probable dans le cas où les traitements de sa cognition sont effectués dans les meilleures conditions. Une interaction parfaite serait alors décrite de la sorte :

« Utilisateur lucide, ayant une perception complète de la situation et de l'outil. Cet utilisateur aura été entraîné à l'utilisation de l'outil pour cette situation. Il sera également disposé cognitivement à l'analyse conceptuelle. Pour mener à bien cette analyse, l'utilisateur aura été instruit et sera informé en temps réel sur les caractéristiques de l'outil, son fonctionnement et l'interprétation de ses communications. »

A l'inverse, une utilisation inappropriée de l'outil serait probable lors d'une interaction décrite ainsi :

« Utilisateur dans un état émotionnel marqué, comme la colère ou l'euphorie, ayant de mauvaises perceptions de la situation, n'ayant jamais utilisé l'outil et n'ayant jamais rencontré cette situation ; dans un état cognitif surchargé réduisant ses capacités d'analyses

conceptuelles. Il n'aura pas été instruit de connaissances explicites concernant les caractéristiques et le comportement de l'outil, qui d'ailleurs ne communiquera pas ses états lors de l'utilisation.»

4.4.7 Variable qualitative pour une interaction en confiance

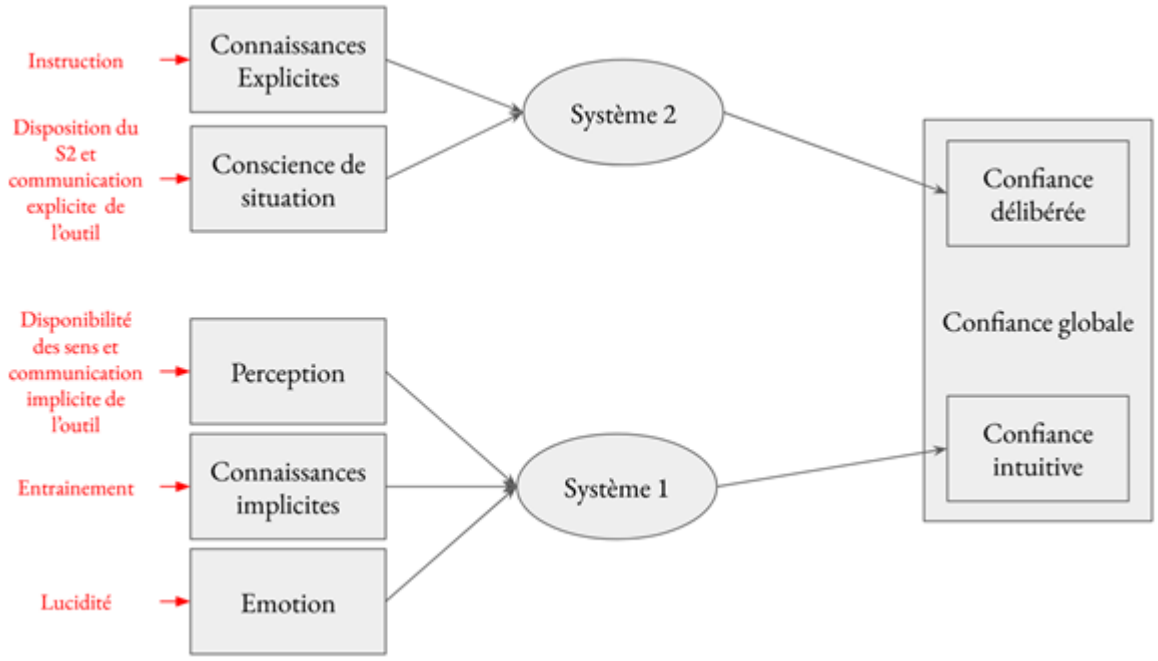

***Figure 4.4.8** : Variables qualitative (en rouge) qui entrent en jeu lors de la génération de la confiance.*

Les variables qualitatives qui, selon notre modèle, déterminent la bonne calibration de la confiance de l'utilisateur sont :

L'instruction. L'instruction est le savoir, composé de connaissances explicites, qui sera utilisé par le système délibéré de l'utilisateur pour juger de la confiance à attribuer à l'outil, dans la situation et pour l'objectif particulier de l'utilisateur. Il s'agit de discours conceptuels établissant des relations de causes à effets entre l'outil, ses caractéristiques, son fonctionnement, la situation et l'objectif de l'utilisateur. Le fait que l'outil soit certifié ou non est par exemple une connaissance explicite que l'utilisateur peut acquérir lors de son instruction. L'utilisateur pourrait également être instruit de connaissances générales sur le fonctionnement, les capacités et limites de la classe de système auquel appartient l'outil. Il serait préférable que l'utilisateur connaisse le fonctionnement exact de son outil mais ce prérequis est difficilement atteignable tant les utilisateurs sont nombreux et les systèmes

d'IA complexes et diversifiés. Dans ce sens, nous avons proposé une classification des systèmes intelligents. Elle a pour objectif de permettre d'instruire les utilisateurs en ce qui concerne les caractéristiques générales du système intelligent qu'il utilise. Une démarche similaire a été proposée par Meta AI avec la mise en place d'un système de cartes instructives sur le fonctionnement d'un système d'IA [Adkins et al. 2022]. Ce type de carte agit comme une pièce identifiant le système et ses données d'entraînement. La carte énonce également, et de manière explicite, les risques liés à son utilisation. D'autres grandes entreprises développant des systèmes d'IA tels que OpenAI ou stability.ia ont depuis adopté cette méthode et publient des cartes d'information lors de la mise sur le marché ou le partage de leur système.

Disposition du système délibéré et communication explicite de l'outil. Lors de l'interaction entre l'utilisateur et le système, il est possible, voire nécessaire, que l'outil informe l'utilisateur d'un ensemble de faits concernant son état et ses productions. Ces connaissances explicites sur la situation, transmises par le système, permettent à l'utilisateur d'améliorer sa conscience de la situation et de mieux décider. Elles concourent aux jugements que porte le système délibéré de l'utilisateur sur la confiance à attribuer à l'outil pour le contexte particulier dans lequel il se trouve. Cependant pour que ces communications de la machine soient utiles, l'utilisateur doit être disposé à les recevoir et à les interpréter conceptuellement pour en comprendre les implications de causes et d'effets. A l'inverse, si l'utilisateur n'est pas disposé cognitivement, comme dans le cas de « l'effet tunnel », il est probable qu'il n'interprète pas, ou n'interprète pas correctement les communications de l'outil. Les explications qui cherchent à être permises par les recherches dans le X.A.I. rentrent dans le cadre des communications explicites de l'outil.

Disposition des sens et communication implicite de l'outil. Les communications implicites de l'outil sont les signaux qui informent implicitement sur l'état et les traitements du système. Il peut s'agir de voyants lumineux, de signaux sonores ou de retours haptiques qui ciblent le sens du toucher. Bien entendu, l'intérêt de ces communications implicites nécessite la disposition sensorielle de l'utilisateur. Cet aspect nécessite l'entraînement de l'individu à l'utilisation de l'outil, afin de s'accoutumer à l'interprétation de ces signaux. Dans leur ensemble, ces communications de la machine, qui peuvent être involontaires comme l'est le bruit d'un moteur par exemple, servent à enrichir les connaissances implicites de l'utilisateur sur l'état de l'outil et de son fonctionnement. Elles aident son système intuitif

à calibrer la confiance situationnelle qui peut être allouée à l'outil en fonction de son expérience.

Entraînement. L'entraînement permet à l'utilisateur de grandir son expérience et de construire des connaissances implicites sur l'utilisation de l'outil pour qu'elles deviennent un savoir-faire. Il s'agit de créer les automatismes de perception et de contrôle qui permettent à l'individu d'utiliser correctement l'outil et d'en interpréter inconsciemment les états, grâce à son expérience. Pour que l'entraînement soit le plus bénéfique il faut, dans la mesure du possible, qu'il représente une simulation proche de la situation d'utilisation.

Lucidité. Chez l'individu humain, les émotions jouent un rôle prépondérant. Elles affectent notre comportement, nos pensées et jugements. Un utilisateur dans un état d'euphorie ou de peur aurait tendance à se retrouver en excès de confiance, respectivement en défaut de confiance. Dans le modèle IPSEL, nous avons suggéré que les émotions influencent les traitements du système intuitif. L'état émotionnel influence donc l'évaluation de la confiance intuitive. Les traitements du système intuitif sont à l'origine du mécanisme de conceptualisation qui nourrit les traitements du système délibératif. Indirectement les émotions influencent aussi l'évaluation de la confiance délibérée. Cependant cette influence sur le système délibéré est moins importante dans le sens où il est possible à l'utilisateur de conscientiser son état émotionnel et ainsi de corriger volontairement les biais que ces émotions pourraient entraîner. Pour cela, encore une fois, l'individu doit être disposé à entamer des réflexions conscientes, ce qui n'est pas toujours le cas, spécifiquement quand il est dans un état émotionnel chargé. Une bonne décision, un bon jugement, une bonne calibration de la confiance attribuable, est plus probable pour un individu lucide, qui n'est pas marqué par une émotion particulière. L'accès à cette lucidité lors de l'utilisation peut sûrement être favorisé par des exercices préalables à l'emploi comme la concentration, la méditation ou le suivi psychologique. Lors de l'utilisation, cette lucidité requise pourrait être mesurée par des moyens de monitoring des états physiologiques de l'utilisateur, ce qui pourrait permettre aux systèmes d'adapter ses communications afin de rétablir une certaine lucidité chez son utilisateur.

4.4.8 Synthèse des discussions

Concernant l'IA de confiance, nous avons dans cet essai présenté un rapide état des lieux des directions prises par la communauté. La majorité des recherches se concentre sur la possibilité d'expliquer conceptuellement la manière dont les algorithmes fonctionnent,

dont les systèmes se comportent (domaine du X.A.I). Cependant de telles explications sont difficiles à acquérir, à représenter et à communiquer. Une seconde caractéristique attendue des systèmes de confiance concerne l'éthique. Nous trouvons que ce sujet est complexe et doit être discuté au niveau sociologique. L'éthique est une construction socio-culturelle, elle est discutable, subjective et contextuelle. Il serait illusoire de croire qu'il est possible de programmer une machine universellement éthique. Elle n'est qu'un outil dont seule l'utilisation peut être qualifiée d'éthique, ou non, relativement aux normes d'une société particulière. Pour collectivement encadrer l'éthique des machines, des normes et règlement sont proposés par diverses institutions.

A la suite de cet état des lieux, nous avons proposé d'utiliser le modèle IPSEL pour identifier les variables qualitatives qui entrent dans la composition d'un système de confiance, ou plus généralement dans une interaction Homme-IA de confiance. Selon nous, une telle interaction nécessite que l'utilisateur humain soit instruit, entraîné, disposé et lucide. Le système quant à lui devrait être certifié et capable de communiquer ses états, ses productions et de potentielles explications, à différents niveaux.

Chapitre 5 : Vers une intelligence générale artificielle

5.1 Proposition d'un pseudo-code

Les chapitres précédents ont présenté un modèle de système de traitement de l'information disposant de mécanismes pour construire sa propre logique à partir de son utilisation, contrôlée ou autonome. L'architecture IPSEL a été décrite et commentée et plusieurs pistes de discussion sur l'intelligence générale artificielle ont été exposées. Il s'agit maintenant de présenter comment un tel système pourrait être codé.

Pour montrer une manière d'implémenter l'architecture globale des processus du modèle IPSEL, nous décrivons ici un programme informatique à la manière d'un pseudo-code. Un pseudo-code s'attache aux mécanismes d'enchaînement des différentes parties du code entre elles. Seules les fonctions et leurs appels sont explicités et non leurs contenus.

5.1.1 La classe CognitionIPSEL : Attributs

La cognition d'un système est un objet informatique. Nous définissons une classe CognitionIPSEL comme suit :

```
57  // La cognition d'un agent est un objet
58  // qui possède des RESSOURCES sous formes d'attributs
59  // et des FONCTIONS sous la formes de méthodes
60  class CognitionIPSEL {
61
62      // l'objet cognition est instancié avec :
63      // - un ensemble de connaissances innées
64      // - une fonction objectif qui retourne l'état de l'agent lorsqu'elle est appellée
65      constructor(innate_knowledge, objectiveFunction) {
66
```

***Figure 5.1.1** : Pseudo-Code, Classe CognitionIPSEL et son constructeur.*

Deux éléments sont définis lors de l'instanciation d'un objet de la classe CognitionIPSEL :

1. un ensemble de connaissances innées
2. une fonction liée à la valuation de la métrique relative à l'atteinte d'un objectif (objective function)

Un objet CognitionIPSEL possède plusieurs attributs qui correspondent aux ressources qu'utilise la cognition de l'agent.

```
65    constructor(innate_knowledge, objectiveFunction) {
66
67      // ------------------------------------------------------------ RESSOURCES de la cognition
68      this.STRUCTURAL_KL = innate_knowledge; // Connaissances innées, non modifiable
69      this.IMPLICIT_KL = []; // Connaissances implicites, modifiables
70      this.EXPLICIT_KL = []; // Connaissances explicites, modifiables
71      this.EMOTION = objectiveFunction; // Etat de l'agent donné par la fonction d'évaluation
72      this.EXPERIENCE = []; // Historique des expériences de l'agent
73      this.EMERGENT_SEQ = []; // Sequénce de représentation émergente
74      this.DELIBERATED_SEQ = []; // Séquence de représentation générée
75
76    }
```

***Figure 5.1.1-2** : Pseudo-Code, Attributs d'un objet CognitionIPSEL.*

Tout d'abord il y a les mémoires, ou ressources, de chaque système de traitement de l'information.

- Les ressources du système Direct (S0), dans notre code STRUCTURAL_KL, qui correspondent aux connaissances innées de l'agent. Ces ressources ne sont pas modifiables car le système Direct n'effectue pas d'apprentissage.

- Les ressources du système Intuitif (S1), dans notre code IMPLICIT_KL, qui correspondent aux connaissances implicites de l'agent. Initialement ces ressources sont vides. Ces ressources sont modifiables pour permettre au système Intuitif (S1) d'apprendre.

- Les ressources du système Délibératif (S2), dans notre code EXPLICIT_KL, qui correspondent aux connaissances explicites de l'agent. Initialement ces ressources sont vides. Ces ressources sont modifiables pour permettre au système Délibératif (S2) d'apprendre.

Ensuite, l'objet CognitionIPSEL possède un attribut représentant son état interne à chaque instant. Dans notre code il s'agit de l'attribut EMOTION. Cet état est évalué par la fonction objectif qui a été définie lors de l'instanciation de la cognition de l'agent.

Enfin, l'objet CognitionIPSEL possède trois attributs utiles pour ses traitements courants :

- un historique de ses expériences passées, dans notre code EXPERIENCE

- un historique des séquences de représentations qui émergent des traitements du système Intuitif (S1), dans notre code EMERGENT_SEQ

- un historique des séquences de représentations générées par le système Délibératif (S2), dans notre code DELIBERATED_SEQ

Le pseudo-code ne précise pas le format de données de ces différents attributs. Ces formats dépendent du problème et de l'environnement pour lesquels l'agent est conçu. Les connaissances implicites pourraient par exemple prendre la forme de matrices numériques alors que les connaissances explicites prendraient la forme d'un multi-graphe d'objets. Ces choix doivent faire l'objet de tests et leur implémentation appartient à la phase de paramétrage. Ces implémentations pourraient être dynamiques, dans ce cas la définition des formats de données de ces différents attributs serait donnée par les connaissances innées passées à l'objet CognitionIPSEL lors de son instanciation.

5.1.2 La classe CognitionIPSEL : Méthodes

L'objet CognitionIPSEL possède également des méthodes qui réalisent les différentes fonctions de la cognition en utilisant les ressources à sa disposition.

Il y a tout d'abord les fonctions sensori-motrices :

```
sensor (stimuli){
    // Transformation des stimulis reçus en messagaes perceptifs
    var percepts = transform(stimuli);

    // appel des traitements du S0 dans THREAD 0
    this.system0(percepts);

    // appel des traitements du S1 dans THREAD 1
    this.system1(percepts);
}

effector (command){
    // Transformation des commandes de controls en actions
    var action = rev_transform(command)

    // Mise en oeuvre des actions
    operate(action);
}
```

***Figure 5.1.2** : Pseudo-Code, Méthodes responsables des fonctions sensori-motrices d'un objet CognitionIPSEL.*

La captation de données provenant de l'environnement externe de l'agent est réalisée par une ou plusieurs fonctions, dans notre code cela est fait par la méthode sensor(stimuli). Cette méthode opère selon la séquence suivante :

1. les stimuli sont transformés en information intelligible pour l'objet cognition. Il s'agit d'un formatage des données réalisé par la fonction transform(stimuli)
2. les stimuli formatés dans un format interprétable par le système sont alors appelés percepts et transmis de manière parallèle aux systèmes Direct (S0) et Intuitif (S1)

Le pseudo-code ne précise pas si c'est exactement le même ensemble de percepts qui est envoyé aux système 0 et 1, ou si ce sont des ensembles distincts. On pourrait également envisager que différentes méthodes sensor(stimuli) soient implémentées et que chacune communique de manière exclusive avec le système 0 ou le système 1. Ces différentes possibilités doivent être testées. Ces choix doivent être faits en fonction du problème et de l'environnement de l'agent et appartiennent à la phase de paramétrage. Ils pourraient également être faits dynamiquement, dans ce cas ces différentes structures et les conditions de leur implémentation seraient données par les connaissances innées passées à l'objet cognition lors de son instanciation.

Les fonctions effectrices de l'agent sont réalisées par une ou plusieurs méthodes, dans notre code ces fonctions sont nommées effector(command). Cette méthode opère selon la séquence suivante :

1. les commandes qui sont reçues des systèmes de traitements sont transformées en actions, un format de données qui est compris par l'organe effecteur devant être mis en mouvement
2. les actions sont mises en œuvre

Le pseudo-code ne précise pas le nombre de fonctions effectrices déployées. Ce choix est fonction du problème et de l'environnement de l'agent et appartient à la phase de paramétrage. Si plusieurs fonctions effectrices sont déployées, leurs opérations sont réalisées en parallèle et leurs appels doivent être faits selon une orchestration particulière pour permettre des comportements coordonnés complexes.

Puis viennent les fonctions correspondantes aux traitements des systèmes 0, 1 et 2 de l'architecture IPSEL.

```
99      // TREAD principale numéro 0
100     system0 (percepts){
101         // Production de commandes motrices à partir des perceptions et des connaissances structurelles
102         var direct_command = reactive_compute(percepts, this.STRUCTURAL_KL);
103
104         // Transmission des commandes motrices aux effecteurs pour mise en oeuvre
105         this.effector(direct_command);
106     }
```

***Figure 5.1.2-2** : Pseudo-Code, Méthode correspondant aux traitements du système Direct (S0) d'un objet CognitionIPSEL.*

Les traitements du système Direct (S0) sont réalisés par une méthode system0(percepts) dans notre code. Cette fonction est mise en œuvre dans un thread principal appelé THREAD 0. Elle prend un ensemble de percepts en entrée et utilise les ressources structurelles STRUCTURAL_KL pour générer des commandes *direct_command* qui seront transmises aux effecteurs.

Ce que le pseudo-code ne précise pas est la manière dont les effecteurs sont appelés. Dans le cas où l'agent possède plusieurs méthodes effectrices, la fonction system0(percepts) pourrait disposer d'une suite d'instructions répartissant les commandes aux effecteurs concernés qui agissent alors en parallèle dans de multiples threads secondaires.

```
108     // TREAD principal numéro 1
109     system1 (percepts){
110         // Production de commandes motrices à partir des perceptions, de l'émotion courante, des productions du S2 et des connaissances implicites
111         // Les traitements font potentiellement émerger une structure qui est placé dans la variable @attention
112         var automatic_command, attention = inductive_compute([percepts, this.EMOTION, this.DELIBERATED_SEQ], this.IMPLICIT_KL);
113
114         // Transmission des commandes motrices aux effecteurs pour mise en oeuvre
115         this.effector(automatic_command);
116
117         // Enregistrement de l'expérience
118         this.EXPERIENCE.push([percepts, automatic_command, this.EMOTION]);
119         this.EMERGENT_SEQ.push(attention);
120
121         // Appel des traitements du S2 dans THREAD 2
122         this.system2();
123     }
```

***Figure 5.1.2-3 :** Pseudo-Code, Méthode correspondant aux traitements du système Intuitif (S1) d'un objet CognitionIPSEL.*

Les traitements du système Intuitif (S1) sont réalisés par une méthode system1(percepts) dans notre code. Cette fonction est mise en œuvre dans un thread principal appelé THREAD 1. Elle prend un ensemble de percepts en entrée et utilise les ressources implicites IMPLICIT_KL, l'état courant de l'agent EMOTION ainsi que les éléments

produits par le S2 enregistrés dans DELIBERATED_SEQ pour générer des commandes *automatic_command* qui seront transmises aux effecteurs. De cette activité opérée sur les ressources de l'agent émerge certaines structures qui seront alors placées dans une variable *attention*.

Le triplet percepts/automatic_command/EMOTION est enregistré comme nouvelle expérience de l'agent dans son historique EXPERIENCE. La structure émergente enregistrée dans la variable *attention* est quant à elle ajoutée à l'attribut EMERGENT_SEQ.

Le système Intuitif (S1) possède également le pouvoir d'appeler la méthode correspondant aux traitements du système Délibératif (S2) dans un troisième thread principal appelé THREAD 2.

Ce que le pseudo-code ne précise pas concerne la ou les conditions d'appel du S2. Cet appel pourrait effectivement être conditionné par la disponibilité des ressources énergétiques de l'agent ou par une évaluation de la confiance attribuée aux *automatic_command* nouvellement générés. Dans ce cas, ces éléments pourraient être encodés dans l'attribut EMOTION, l'état de l'agent conditionnera alors sa propension à faire appel à son système délibératif. Le pseudo-code ne précise pas non plus la manière d'identifier les éléments émergents qui doivent être placés dans la variable attention.

```
126     // THREAD principal numéro 2
127     system2 (){
128        // Production d'une séquence de représentation à partir de la séquence émergente et des connaissances explicites
129        this.DELIBERATED_SEQ.push(deductive_compute(this.EMERGENT_SEQ, this.EXPLICIT_KL));
130     }
```

Figure 5.1.2-4 : *Pseudo-Code, Méthode correspondant aux traitements du système Délibératif (S2) d'un objet CognitionIPSEL.*

Les traitements du système Délibératif (S2) sont réalisés par une méthode system2() dans notre code. Cette fonction est mise en œuvre dans un thread principal appelé THREAD 2. Elle utilise la variable globale représentant l'attribut EMERGENT_SEQ et les connaissances explicites EXPLICIT_KL pour générer des séquences de représentations explicites qui sont ajoutées à l'attribut DELIBERATED_SEQ. Puisque cette ressource DELIBERATED_SEQ est utilisée par le système Intuitif (S1), les traitements du système Délibératif (S2) ont bien le potentiel d'influencer et de moduler les commandes automatiques produites par le S1.

Ce que le pseudo-code ne précise pas concerne le nombre de représentations explicites qui sont ajoutées à la liste DELIBERATED_SEQ. Il serait possible d'ajouter les représentations une par une, ou bien d'ajouter des séquences de représentations. Ce choix devrait être testé et appartient à la phase de paramétrage.

```
132     learn_implicit (){
133         // Construction et modification des connaissances implicites à partir de l'expérience
134         update_model_implicit(this.EXPERIENCE, this.IMPLICIT_KL);
135     }
136
137     learn_explicit (){
138         // Construction et modification des connaissances explicites à partir des connaissances implicites
139         update_model_explicit(this.IMPLICIT_KL, this.EXPLICIT_KL);
140     }
```

Figure 5.1.2-5 *: Pseudo-Code, Méthode correspondant aux mécanismes d'apprentissage d'un objet CognitionIPSEL.*

Enfin, deux méthodes de la classe CognitionIPSEL permettent à l'objet de modifier ses ressources de manière autonome.

La méthode learn_implicit() utilise l'expérience de l'agent, accessible via l'attribut EXPERIENCE pour mettre à jour les connaissances implicites. Il s'agit pour cette méthode de généraliser l'expérience de l'agent à la manière d'une induction. Pour rappel, EXPERIENCE contient une liste de triplets (Percepts / Automatic_command / EMOTION) représentant les expériences passées de l'agent. Cette méthode a donc pour objectif de trouver les liens récurrents entre perceptions, actions et état de l'agent et de les représenter en tant que connaissances implicites qui seront utilisées par le système Intuitif (S1). De ce fait, pour un ensemble de perceptions données, le système Intuitif utilise les connaissances implicites pour trouver les actions qui engendreront les états émotionnels les plus désirables.

La méthode learn_explicit() utilise les connaissances implicites pour constituer et modifier les connaissances explicites qui seront utilisées par le système Délibératif (S2). Il s'agit de trouver des structures et des motifs récurrents dans l'ensemble des connaissances implicites, ainsi que des relations dynamiques entre ces différentes formes. Ainsi, le rôle des DELIBERATED_SEQ est de fournir une source d'information auto-générée qui a le potentiel de représenter les dynamiques entres perceptions, action et émotion, permettant ainsi de prédire les évolutions futures de l'agent et de l'environnement pour mettre en œuvre

des actions qui ne sont pas simplement commandées automatiquement mais qui sont également délibérées.

Ce que le pseudo–code ne précise pas concerne les moments où ces deux fonctions d'apprentissage sont appelées. Elles pourraient être appelées en continue au cours de la vie de l'agent, comme elles pourraient n'être appelées qu'à certains moments, en fonction de conditions relatives à l'état courant de l'agent.

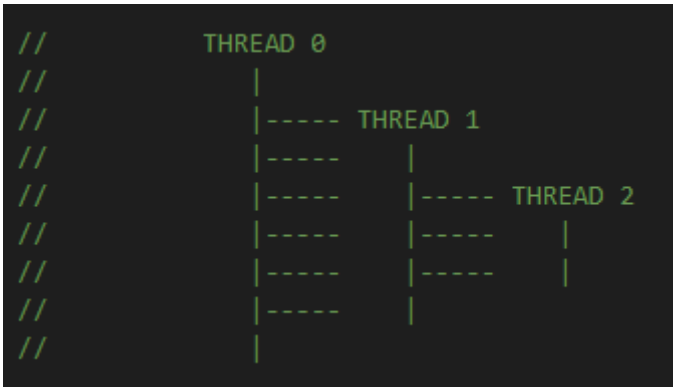

Figure 5.1.2-6 : *Multi-threading de l'objet CognitionIPSEL*

Pour conclure sur la présentation du pseudo-code d'un objet implémentant l'architecture IPSEL, nous insistons sur le parallélisme de ses traitements. Chaque système de traitement de l'information possède son propre thread principal et tous utilisent des ressources communes implémentées en tant que attributs de l'agent. Le système Direct fonctionne en continu et utilise des ressources présentes dès l'instanciation de l'agent en tant que connaissances innées qui ne sont pas modifiables, assurant ainsi la pérennité et l'indépendance des traitements du S0. Le système Intuitif fonctionne également en continu, mais il utilise des ressources qui doivent en premier lieu être construites à partir des suites de l'apprentissage implicite de l'agent. Dépendantes de l'expérience et des émotions de l'agent, les ressources implicites permettent à l'agent de s'adapter à son environnement en mobilisant des connaissances induites de ses expériences passées. Enfin le système Délibératif ne fonctionne pas en continu, le recours à ses traitements est conditionné par l'appel qui en est fait par le système Intuitif et l'état des ressources explicites qui doivent également d'abord être construites.

Implémenter un agent dont la cognition suit les principes de l'architecture IPSEL reviendrait à mettre en œuvre le pseudo-code que nous venons de présenter et qui est

retranscrit dans son intégralité en annexe 3 de ce document. Pour autant de nombreux choix d'implémentation sont à considérer pour arriver à un fonctionnement effectif. En fonction de l'environnement et de la problématique que l'agent doit y résoudre, il faut : spécifier et implémenter les connaissances innées et la fonction objectif de l'agent, définir les fonctions sensori-motrices par rapport aux modalités perceptives et motrices voulues, choisir les formats de données des différentes ressources de l'agent, définir les fonctions d'apprentissage par rapport aux formats de données choisis et enfin procéder à de multiple tests de paramétrage voire à un procédé évolutionniste pour les optimiser.

5.1.3 Pistes d'implémentations

Dans les deux parties précédentes nous avons décrit un pseudo-code qui montre une manière de concevoir le squelette global de l'architecture IPSEL. Il reste néanmoins beaucoup de choses à spécifier pour arriver à une implémentation fonctionnelle : les formats de données, le contenu des fonctions et les méthodes d'apprentissages. Sur ces points plusieurs options sont possibles et envisageables, de la même manière plusieurs paramétrages sont à tester pour envisager d'atteindre les fonctionnalités recherchées.

Les choix architecturaux primordiaux qui, selon nous, permettent à l'architecture IPSEL de mettre en œuvre une intelligence générale artificielle sont : l'intrication de deux niveaux de traitement de l'information, implicite et explicite, et la possibilité pour l'agent de construire ses propres connaissances pour modéliser l'environnement avec lequel il est en interaction. Dans cette partie nous proposons une discussion des moyens techniques contemporains qui pourraient permettre d'implémenter informatiquement ces capacités.

Si nous devions concevoir un tel agent, voici les directions techniques que nous explorerions :

Format des données

Dans le premier niveau de traitement, l'information prend une forme implicite, elle représente l'état des organes senseurs, les états de l'agent que nous avons appelé *Émotions,* l'état des organes effecteurs et les connaissances implicites. Ces informations pourraient prendre la forme de valeurs numériques, vecteurs de valeurs numériques ou matrices de valeurs numériques. Pour simplifier notre description, nous regroupons l'ensemble de ces formats de données sous l'appellation matrice. Les traitements du système intuitif (S1) sont

alors des calculs matriciels. Les résultats informatiques des communautés de l'intelligence artificielle montrent que ce type de calcul s'avère très efficace pour réaliser les fonctions que nous avons attribuées au système intuitif (S1), principalement les fonctions de perceptions, de reconnaissances et de contrôle moteur.

Au deuxième niveau de l'architecture IPSEL, les traitements sont effectués sur de l'information représentée sous une forme explicite. Ces représentations sont construites à partir des représentations implicites et cela permet l'intrication entre les deux niveaux. Il s'agit de construire des entités explicites, des ensembles de règles, à partir des matrices utilisées pour représenter les connaissances implicites. Les règles peuvent porter sur la définition des entités et aussi sur leurs relations avec d'autres entités explicites.

L'ensemble des représentations explicites pourrait alors être représenté par un hypergraphe multidimensionnel qui organise l'information en tant qu'entités (les sommets du graphe), ayant des attributs (caractéristiques des sommets du graphe) et des relations (les arêtes du graphes). De cette manière, l'état du monde courant, les connaissances portant sur sa dynamique et des représentations d'état simulés du monde pourraient être explicitement représentés par des graphes. Les traitements du système délibératif (S2) consisteraient alors à parcourir ces graphes pour trouver des chemins et des sous-graphes particuliers. Les résultats informatiques des communautés de l'intelligence artificielle montrent que ce type de manipulation de graphe s'avère très efficace pour réaliser les fonctions que nous avons attribuées au système délibératif (S2) comme l'analyse et le raisonnement logique.

Pour résumer la piste que nous préconisons d'explorer en premier, le système intuitif (S1) représente les données sous la forme de matrice et ses traitements reposent sur le calcul matriciel, alors que le système délibératif (S2) représente les données sous la forme de graphe et ses traitements reposent sur la manipulation de graphe.

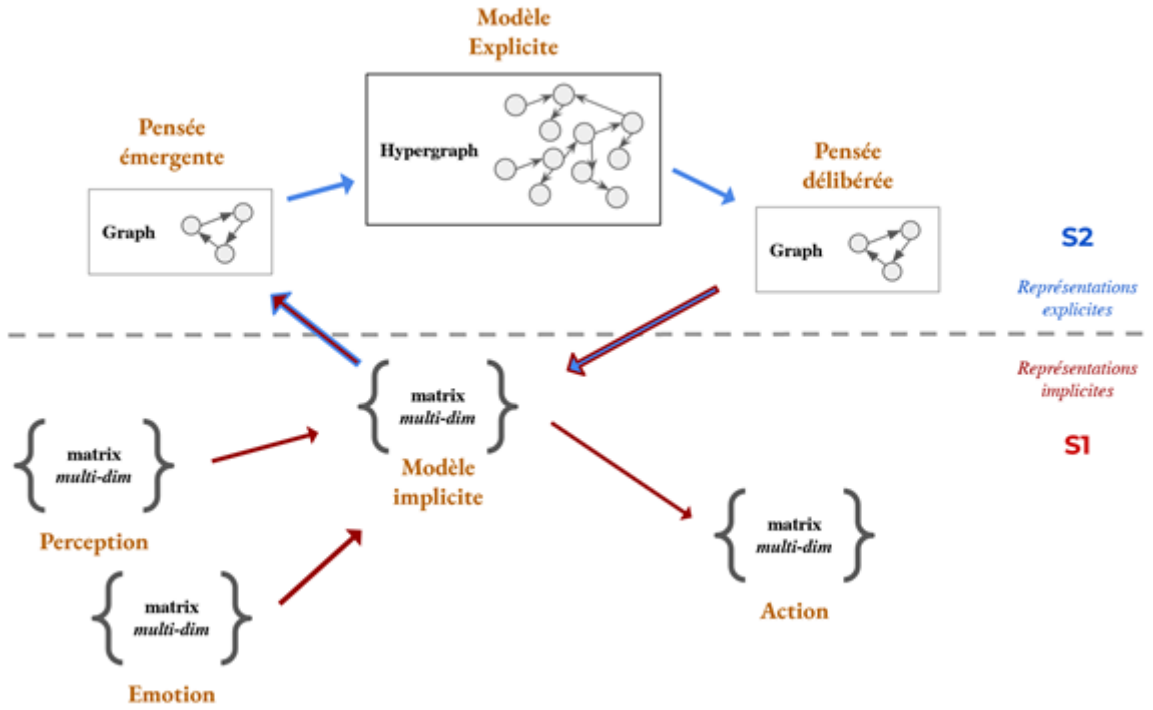

Figure 5.2.6 : *Exemple de format et flux de données transitant entre les différents modules fonctionnels de l'architecture IPSEL.*

Fonctions

Les différents formats de données subissent des transformations qui constituent les traitements du système. Une liste de ces transformations est donnée dans le tableau de la figure suivante :

Transformation		Fonction de la transformation
Perception *Matrice*	**Modèle implicite** *Matrice*	**Représentation implicite de l'état de l'environnement externe** *(Sentir)*
Emotion *Matrice*	**Modèle implicite** *Matrice*	**Représentation implicite de l'environnement interne** *(Ressentir)*
Modèle implicite *Matrice*	**Action** *Matrice*	**Contrôle implicite des effecteurs** *(Agir)*
Modèle implicite *Matrice*	**Pensée émergente** *Graph*	**Représentation explicite de l'état du monde** (monde = env. externe + env. interne) *(Représenter explicitement l'état implicite, conscientiser)*
Pensée émergente *Graph*	**Modèle explicite** *Graph*	**Trois fonctions de délibération** *(Réfléchir)* **:** 1. Cohérence du graph avec les connaissances explicites *(Comprendre explicitement)* 2. Abduction des causes du graph en fonction des connaissances explicites *(Expliquer)* 3. Déduction des effets du graph en fonction des connaissances explicites *(Prédire)*
Modèle explicite *Graph*	**Pensée délibérée** *Graph*	**Résultat des délibérations**
Pensée délibérée *Graph*	**Modèle Implicite** *Matrice*	**Objectivation de la pensée** *(Intégrer implicitement le résultat des réflexions)*

Figure 5.2.6-2 : *Tableau établissant les transformations des formats de données et leurs fonctions.*

La complémentarité entre un niveau implicite et un niveau explicite est étudiée dans la discipline nommée neuro-symbolisme [Lemos 2020]. On peut également la voir comme la combinaison des approches connexionnistes et cognitivistes.

La construction des connaissances de l'agent est un problème d'apprentissage qui prend différentes formes en fonction de ce qui est à apprendre.

Nous distinguons 4 mécanismes d'apprentissage : (1) Construire les connaissances implicites (en bleu sur le schéma 5.2.6-3) ; (2) Construire les connaissances explicites à partir des connaissances implicites (en vert sur le schéma 5.2.6-3) ; (3) Manipuler efficacement les connaissances explicites (en orange sur le schéma 5.2.6-3) ; (4) Calibrer les émotions de telle sorte qu'elles dirigent correctement l'agent vers la réalisation de ses objectifs (en rose sur le schéma 5.2.6-3).

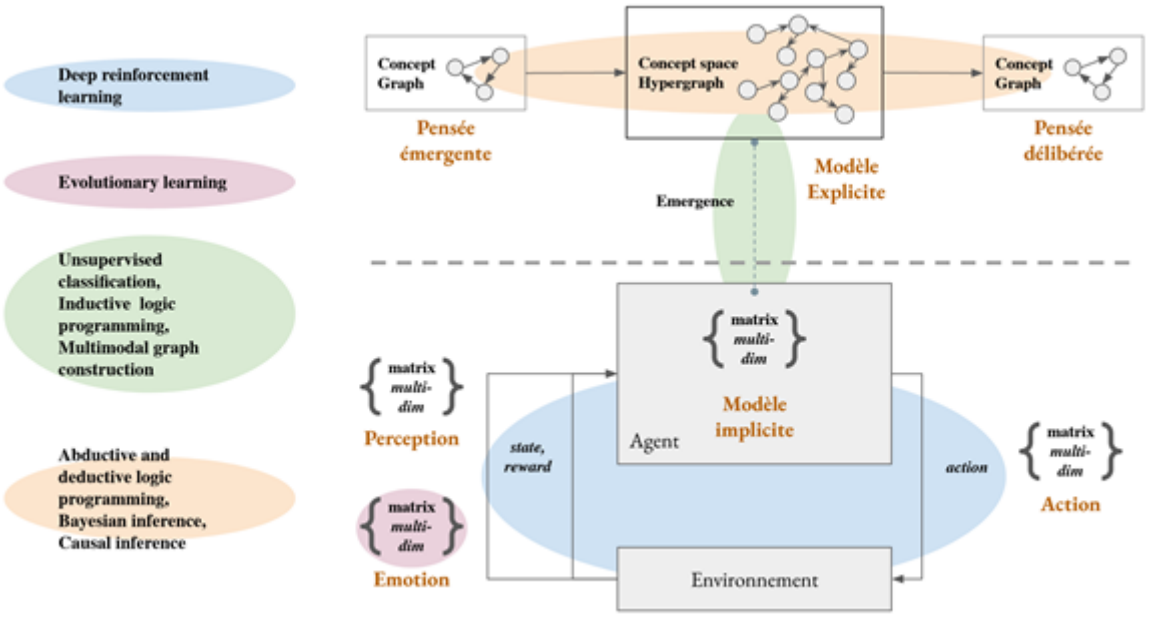

Figure 5.2.6-3 : *Les différents champs d'études de l'IA qui explorent des techniques algorithmiques qui permettraient de résoudre les différents problèmes d'apprentissage de l'architecture IPSEL.*

(1) La construction des connaissances implicites équivaut à l'apprentissage d'un modèle. Ce modèle permet à l'agent de trouver, en fonction de l'état courant des perceptions et celui des émotions, les commandes à mettre en œuvre. Les données manipulées lors de ce traitement sont des valeurs numériques regroupées sous forme de matrices. Les technologies du « Machine learning » telles que les réseaux de neurones effectuent ce type de calcul matriciel et apprennent à construire de tels modèles implicites. Plusieurs architectures de réseaux pourraient être testées. En particulier les architectures du type Transformer semblent être une direction de premier choix compte tenu de la généralité de leurs capacités de représentations (les réseaux de neurones à convolution semblent appropriés pour représenter les relations spatiales, et les réseaux de neurones à mémoire courte et longue pour représenter les relations temporelles). Néanmoins, le domaine des réseaux de neurones évolue encore beaucoup et il ne serait pas surprenant que de nouvelles architectures entremêlant des principes provenant de différentes structures de réseaux neuronaux voient le jour. En outre, les valeurs de la ou des matrices représentant le modèle implicite sont apprises sur un mode essai-erreur qui peut être formalisé par un problème d'apprentissage par renforcement (reinforcement learning)

[Kaelbling 1996] où l'agent apprend de manière autonome [Sharma et al. 2021]. Ce processus correspond aux technologies d'apprentissage par renforcement profond (Deep Reinforcement Learning) [Arulkumaran 2017][Chen 2021]. Une technique qui fait actuellement beaucoup parler d'elle par les résultats qu'elle permet d'atteindre dans de nombreuses problématiques d'intelligence artificielle [Mnih 2015] [Otto 2021].

(2) Construire efficacement les représentations explicites à partir des représentations implicites est un problème de représentation. Spécifiquement il s'agit d'une problématique de classification non supervisée, ou « clustering », où l'objectif est de construire de bonnes représentations [Ronen 2022]. Comme beaucoup de disciplines de l'IA, ce domaine de recherche est en pleine effervescence et de nouvelles techniques semblent particulièrement adaptées aux capacités générales que nous cherchons à donner à notre agent [Higgins 2022].

Les représentations ainsi formées ont vocation à former un hypergraphe à plusieurs dimensions. Cet aspect est également bien étudié par la communauté de l'IA [Zhu 2022]. Un graphe peut être représenté sous la forme d'un ensemble de règles, à condition que les entités du graphe soient définies. La construction des règles, ou dit autrement des relations, liant les entités du graphe équivaut à un problème d'induction logique [Hamilton 2017]. De manière plus générale, la construction de graphe à partir de réseaux de neurones est étudiée par une discipline récente mais qui ne cesse de prendre de l'ampleur : les « Graph Neural Network » [Wu et al. 2020].

(3) Manipuler efficacement les connaissances explicites revient à effectuer différentes opérations sur des graphes comme la vérification de cohérence entre deux graphes, la navigation sous contrainte dans un graphe ou la découverte des plus proches sommets voisins d'un sommet donné. La navigation dans un graphe de connaissances est, en soi, un sujet de recherche prolifique [Zhang & Yao 2022] [Rossi et al. 2021][Ji et al. 2021]. Les graphes pouvant être représentés par des ensembles de règles, trouver dans l'hyper graphe un sous graphe comme cause d'un graphe donné peut prendre la forme d'un problème d'abduction logique [Kakas 1992] (abductive logic programming). De la même manière, trouver dans l'hyper graphe, un sous graphe comme effet d'un graphe donné est un problème de déduction logique (deductive logic programming). Il s'agit d'utiliser des ensembles de règles pour inférer des résultats comme cela est fait par les techniques de raisonnement causal ou bayésien [Pearl 1998] [Bengio 2022].

(4) Calibrer les émotions de telle sorte qu'elles dirigent efficacement l'agent dans la réalisation de ses objectifs est un problème de paramétrage. Plusieurs techniques algorithmiques permettent de résoudre ce type problème mais il semble que celles qui imitent les mécanismes de la sélection naturelle soient actuellement, de notre point de vue, les plus efficaces et prometteuses [Arulkumaran 2019] [Tang et al. 2022] [Liu 2022].

Un ensemble de travaux réalisés par Jürgen Schmidhuber et ses équipes dans les années 90 [Schmidhuber 1991]. Cette équipe de recherche américaine a exploré différentes implémentations de réseaux de neurones artificiels ayant vocation à réaliser un fonctionnement similaire à celui qui est attendu dans notre proposition. Notamment le couplage de deux systèmes de traitement, nommés « controller and world model » par Schmidhuber, et un comportement guidé par un signal d'évaluation interne [Schmidhuber 2015]. La principale différence entre la proposition de Schmidhuber et la nôtre est que, dans la modélisation que nous avons proposée, le système délibératif, respectivement le world model pour Schmidhuber, ne communique pas directement avec les organes effecteurs, alors que cela est le cas dans l'architectures de Schmidhuber.

5.2 Discussions à propos de l'intelligence générale artificielle

La discipline de l'intelligence artificielle est née dans les années 50, par des actes précurseurs comme la réflexion du mathématicien Alan Turing qui se demandait si les machines pouvaient penser [Turing 1950].

Cette possibilité de créer des machines et systèmes dont l'intelligence s'approcherait de celle des individus anime depuis lors les recherches en IA, elle constitue un horizon, une borne supérieure dont la faisabilité, faute de preuves empiriques, est encore débattue à ce jour.

En informatique, Shane Legg et Marcus Hutter ont proposé une définition formelle de l'intelligence pour un agent artificiel. Pour ces chercheurs, l'intelligence mesure la capacité d'un agent à réaliser ses objectifs dans une multitude d'environnement [Legg & Hutter 2007-2]. L'intelligence est alors une valeur calculée. Plus l'ensemble des environnements dans lequel l'agent est capable de réaliser ses objectifs est grand, plus son intelligence est grande. La formule mathématique proposée par Legg & Hutter pour calculer cette mesure de l'intelligence est difficile à appliquer, voire incalculable dans la plupart des cas, mais elle a

le mérite d'insister sur un point qui semble fondamental pour l'intelligence : l'autonomie de l'agent et son adaptation à de multiples domaines.

Cette vision définissant l'intelligence comme une capacité à s'adapter de manière autonome nous apparaît pertinente, malgré son minimalisme, car elle permet d'intégrer les différentes facettes que l'on associe usuellement à l'intelligence. Par exemple, l'intelligence relationnelle et empathique pourrait être considérée comme une capacité de l'agent à adapter son comportement à son interlocuteur ; la créativité comme une faculté permettant de trouver de nouvelles solutions d'adaptation; l'intelligence orale comme une compétence servant l'adaptation à l'environnement social ; etc.

En prenant plus de hauteur et en gardant ce cadre agent-environnement, nous pourrions suggérer qu'en termes de traitements de l'information, l'intelligence vient à s'exprimer quand il est permis à l'agent de:

(1) percevoir et représenter l'état de l'environnement

(2) manipuler les connaissances qui permettent de s'y adapter

(3) contrôler son corps pour y agir de manière autonome.

Ainsi en reprenant la définition de Shane Legg et Marcus Hutter, la mesure de l'intelligence d'un agent est proportionnelle à la taille de l'ensemble des éléments de l'environnement que l'agent peut percevoir et représenter. Elle est également proportionnelle à la taille de l'ensemble des connaissances que l'agent peut manipuler. Et enfin, elle est proportionnelle à l'ensemble des actions que l'agent peut mettre en œuvre de manière autonome.

Il semble que ce soit cette manière de considérer la mesure de l'intelligence qui est utilisée pour distinguer des degrés d'intelligence artificielle : l'intelligence artificielle restreinte (« Artificial Narrow Intelligence (ANI) »), l'intelligence générale artificielle (« Artificial General Intelligence (AGI) ») et l'intelligence artificielle de niveau humain (« Human Level Artificial Intelligence (HL-AI) »).

Un système doté d'une intelligence restreinte (ANI) est en capacité de s'adapter et d'améliorer ses performances pour la réalisation d'une unique tâche, pour un domaine restreint. Il dispose d'un ensemble de connaissances restreint et s'adapte à peu

d'environnement, en l'occurrence uniquement à des environnements spécifiques et bien définis lors de la conception du système.

Un système doté d'une intelligence générale (AGI) est, à l'inverse, un système capable d'apprendre à améliorer ses performances dans la réalisation de multiples tâches opérées dans des domaines différents. L'agent possède de fait un plus grand ensemble de connaissances que s'il ne pouvait opérer que dans un seul domaine. Certaines de ces connaissances sont qualifiées de générales car elles sont utilisables dans plusieurs domaines. Cette généralisation des connaissances permet théoriquement à l'agent de s'adapter à un plus grand ensemble d'environnements ou à des environnements plus complexes, les systèmes dotés d'AGI sont tacitement considérés comme « plus intelligents » que les systèmes dotés de ANI.

Un système doté d'une intelligence artificielle de niveau humain (HL-AI) est un système capable de s'adapter à des environnements réels. Qui plus est, le système met en œuvre les mêmes fonctions que celles de l'intelligence humaine : le raisonnement conceptuel, l'introspection, l'empathie, la compréhension du langage etc.

Ces définitions ne semblent pas formalisables et peuvent paraître ambiguës. L'attribution du caractère d'intelligence à un système sera toujours issue du jugement d'un tiers et donc nécessairement subjective, car dépendante des choix des tâches et des domaines que ce tiers pose comme cadre de son jugement. Par exemple, nous pourrions dire que l'Homme est doté d'une intelligence restreinte si l'on considère qu'il lui est permis de réaliser une unique tâche, survivre, dans un unique environnement, la réalité. Autre exemple, le programme DeepBlue était considéré par tous comme intelligent à la fin des années 90, lorsqu'il a battu le champion d'échecs Gary Gasparov, mais ce programme n'est même plus qualifié d'intelligence restreinte de nos jours. Malgré la subjectivité de l'évaluation que l'on porte sur elles, ces trois types d'intelligences sont utiles pour appréhender les développements technologiques des systèmes modernes. Il est consensuellement admis que les systèmes d'IA actuels peuvent être qualifiés d'intelligences restreintes, eu égard à leurs nombreuses réalisations pour des tâches spécifiques mais à leur incapacité à généraliser leurs connaissances. Pour qu'ils s'intègrent mieux aux activités humaines et atteignent de nouvelles sphères de performances, il est nécessaire de chercher à les doter d'une intelligence plus générale. Faisons remarquer que chercher à construire des intelligences générales artificielles ne signifie pas construire des humains artificiels, car il est surement possible de

concevoir des systèmes capables de généraliser leurs connaissances à travers un ensemble de domaines sans nécessairement avoir recours à des fonctions de l'intelligence humaine, telles que la pensée subjective ou la compréhension du langage naturel.

De nombreuses capacités cognitives qui ont été observées et formalisées chez l'humain manquent aux systèmes dotés d'IA [Marcus 2020]. Ces capacités sont différemment listées et plusieurs auteurs se sont attachés à en trouver des formulations accessibles. Nous prendrons la liste des sept points proposés par une équipe de recherche d'IBM [Ganapini et al. 2022] : (1) L'apprentissage et l'intégration continue de connaissances implicites et explicites, (2) L'abstraction et la généralisation des connaissances, (3) Le raisonnement et l'analyse causale, (4) L'adaptabilité à divers environnements, (5) Le contrôle de soi (self-control), (6) Le sens-commun, (7) Le raisonnement éthique.

L'architecture IPSEL présente le fonctionnement d'un agent dont les comportements sont guidés par des représentations internes de son propre état et de celui de son environnement. Ces représentations internes partagent des relations les unes avec les autres et forment des connaissances pour l'agent qui les manipule. Les connaissances de l'agent s'expriment dans différents degrés d'abstractions. Pour simplifier la modélisation, nous avons établi deux catégories de connaissances : les connaissances implicites, construites par l'agent à partir de son expérience sensori-motrice et utilisées par son système intuitif (S1) ; et les connaissances explicites, bien plus abstraites, construites à partir des connaissances implicites et utilisées par le système délibératif (S2). Les rapports qui se constituent entre expérience sensori-motrice, connaissances implicites et connaissances explicites permettent aux différents systèmes qui s'y réfèrent (organes sensori-moteur, système 1 et système 2) de coopérer entre eux. Nous allons maintenant décrire comment ces principes architecturaux pourraient permettre de mettre en œuvre les différentes capacités humaines qui manquent aux machines et qui ont été listées dans le paragraphe précédent.

(1) <u>L'apprentissage et l'intégration de connaissances implicites et explicites.</u>

Cette capacité est au cœur de l'architecture IPSEL et constitue l'une de ses principales caractéristiques différenciantes, elle résulte de l'intégration de la « dual-process theory » pour un système cognitif artificiel. Les connaissances implicites sont apprises et intégrées en continue par l'agent au cours de ses interactions avec l'environnement. Les connaissances explicites sont apprises et intégrées par l'agent en tant que constructions qui émergent de l'auto-organisation des connaissances implicites. Les connaissances implicites sont

manipulées par le système intuitif (S1) alors que les connaissances explicites sont utilisées par le système délibératif (S2). La relation d'enracinement et de composition entre connaissances implicites et explicites permet la communication, et donc la coopération, des deux systèmes de traitement de l'information S1 et S2.

(2) L'abstraction et la généralisation des connaissances.

La capacité d'abstraction et de généralisation des connaissances est permise dans l'architecture IPSEL par le mécanisme que nous avons nommé « logique émergente ». Ce mécanisme est à l'œuvre lors de la construction des connaissances explicites à partir de l'organisation des connaissances implicites. Les représentations conceptuelles que nous avons décrites dans notre modélisation sont des représentations explicites. Ce sont des abstractions de compositions particulières de représentations implicites. Chez IPSEL, c'est grâce à l'abstraction que les connaissances peuvent être généralisées. Une représentation abstraite est une représentation de certains éléments communs d'un ensemble de représentations de plus bas niveau. La représentation abstraite se détache des spécificités des représentations qu'elle généralise pour n'en garder que certaines caractéristiques saillantes qui sont potentiellement applicables à d'autres domaines que ceux à partir desquelles elles ont été apprises.

(3) Le raisonnement et l'analyse causale

Dans le modèle IPSEL, les connaissances explicites sont construites à partir des connaissances implicites. La différence entre ces deux formes de représentations est que les connaissances explicites sont représentées par leurs aspects interrelationnels. Les représentations explicites nommées concepts sont des représentations des fonctions et du dynamisme associable aux connaissances implicites qui ont permis de les construire et/ou qui peuvent les instancier. En ce sens, un concept peut être explicité en donnant une description de ses relations, là où une représentation implicite ne peut être décrite que par les valeurs d'activation des organes sensoriels dont elle représente l'état. C'est en outre ce qui rend les représentations explicites communicables, et les représentations implicites non communicables puisqu'elles représentent uniquement des états du corps particulier de l'agent, et n'ont donc de sens que pour ce dernier. Par exemple, il est possible d'expliciter le concept « violet » par ses relations à d'autres concepts comme: un mélange entre le rouge et le bleu, ou sa longueur d'onde. Mais il n'est pas possible pour un agent de transmettre la sensation, ce qu'il voit, lorsqu'il perçoit du violet, car cela est totalement relatif à son corps

et plus particulièrement à ses organes visuels. Ainsi, bien que les humains s'accordent sur la définition conceptuelle du violet, il est impossible de savoir si tous les humains voient le même violet, si tous les humains associent ce concept avec les mêmes activités sensorielles car ces activités sensorielles sont spécifiques à leur corps.

Les représentations conceptuelles n'existent donc que par leurs relations avec d'autres représentations, c'est ce qui les rend explicites. Ces relations ont un sens et représentent des rapports de compositions et de causalité. L'espace de concept permet à l'agent de se représenter, de manière abstraite, le monde dans lequel il se meut. Les relations entre les concepts permettent à l'agent de se représenter la dynamique de ce monde, c'est-à-dire les liens de composition et de causalité entre les objets dont les concepts sont l'abstraction. Le raisonnement et l'analyse causale sont alors des capacités permises chez IPSEL par les traitements du système délibératif (S2), qui mobilisent et manipulent les connaissances explicites, liées entre elles par des relations de compositions, de causes et d'effets. Ces manipulations de représentations explicites en utilisant leurs relations les unes aux autres pourraient être considérées comme une capacité de raisonnement et d'analyse causale artificielle.

(4) L'adaptabilité à divers environnements

Pour pouvoir s'adapter à divers environnements, un agent doit : soit avoir des connaissances utiles relatives à chacun des environnements auquel il doit s'adapter, soit être capable d'acquérir ces connaissances par lui-même au cours de son interaction avec chaque environnement. Pour les environnements réels, l'aspect combinatoire, étendu et imprévisible des situations ne permet pas d'espérer pouvoir fournir toutes les connaissances nécessaires à l'agent lors de sa construction. Un agent IPSEL a pour vocation d'acquérir ces connaissances lors de ses interactions avec l'environnement. La capacité d'un agent IPSEL à construire des connaissances explicites, abstraites et généralisables, lui permet alors théoriquement de s'adapter à divers environnements.

(5) Le self control

Le self-control est permis chez IPSEL par l'architecture à deux niveaux de ses traitements de l'information. Au premier niveau, les traitements de l'information s'effectuent d'une manière que l'on a décrit comme automatique en utilisant des connaissances implicites. Ce sont les traitements du système intuitif (S1). Il s'agit de traitements qui déterminent les

actions en fonction des connaissances implicites appliquées aux perceptions, dans un mode qui s'apparente à la réaction ou aux comportements intuitifs des Hommes et des animaux. Dans ce mode d'action que l'on pourrait apparenter à un instinct artificiel, l'agent n'a pas à proprement parler de « self control », il réagit à ses perceptions en utilisant des modèles implicites, que l'on pourrait aussi appeler heuristiques. Cependant, dans l'architecture IPSEL, lorsque le deuxième niveau s'est construit comme émergence provenant de l'auto-organisation des connaissances du premier niveau, un autre chemin de traitement de l'information est possible. Ce second mode de traitement correspond à l'activité du système délibératif (S2) qui effectue ce qui pourrait s'apparenter à un raisonnement artificiel. Les productions de ce système 2, le système délibératif, ont le potentiel d'influencer, de moduler voire d'inhiber les traitements du premier niveau. Cette action du système 2 sur le système 1 peut être, selon nous, considérée comme le self-control de l'agent IPSEL. Par un acte de manipulation de connaissances explicites, le système 2 peut, dans une certaine mesure, contrôler les comportements du système 1 et ainsi moduler les comportements que nous avons associé à une sorte d'instinct artificiel.

(6) Le sens commun

Dans la modélisation IPSEL, le sens commun est l'ensemble des connaissances implicites que l'agent a construites au cours de son expérience. Dans le langage courant et pour l'Homme, l'aspect « commun » de ces connaissances fait référence au fait qu'elles sont communes à la plupart des individus. Pour des agents IPSEL, étant donné qu'ils construisent individuellement leurs propres connaissances, plusieurs agents IPSEL ayant les mêmes modalités sensori-motrices et le même environnement, auraient de grandes chances de construire des connaissances implicites que l'on pourrait qualifier de communes. Elles n'en seraient pas moins individuelles. Il ne semble pas correct de considérer que tous les humains possèdent exactement le même ensemble de connaissances communes, le même sens-commun. Ce qui est du sens commun pour un Japonais ne l'est pas forcément pour un américain, cela est relatif à leurs environnements et aux constructions socio-culturelles qui en découlent. Pour autant, tous les Hommes vivent sur Terre et ont affaire, dès leur naissance, à la gravité, et tous les Hommes ont une compréhension implicite de cette dernière. Pour l'humain aussi, le sens commun est quelque chose d'individuel bien que de nombreuses connaissances implicites soient communes à la plupart des individus. De la même manière, un agent IPSEL développe son propre sens commun, qui pourrait être plus

ou moins proche de celui des humains en fonction de la similarité de leurs expériences respectives.

(7) Le raisonnement éthique

L'éthique comme ensemble de normes encadrant les comportements sociaux d'une communauté peut être apprise et raisonnée par un individu à partir du moment où ce dernier est capable de se représenter conceptuellement en tant qu'individu appartenant à la communauté. Pour un agent IPSEL, le raisonnement éthique est permis par sa capacité à construire une représentation conceptuelle et explicite du monde. Cette représentation du monde peut inclure une représentation de soi et d'autres agents comme conceptuellement équivalent au soi, c'est-à -dire ayant un corps, une volonté et des émotions. Dans cette représentation explicite du monde se sont également construites des connaissances sur les « bonnes manières » d'interagir avec d'autres individus. L'aspect « bon » de ces manières d'interagir peut-être évalué au regard des effets qu'elles produisent pour l'agent, ou par rapport à des normes construites collectivement. Dans ce second cas où les modes d'interactions sont évalués par rapport à une norme socio-culturelle, on pourrait les considérer comme suivant une certaine éthique.

Dans son article fondateur, Alan Turing a discuté de l'objection qui consiste à dire qu'il ne sera jamais possible de doter les machines de telle ou telle capacités. Une objection générale qui prendrait la forme « Vous ne serez jamais capable de construire une machine qui fasse X » où X est une capacité que l'on suppose inaccessible aux machines [Turing 1950]. Il faisait alors remarquer que ce type de présupposé n'était jamais démontré ou prouvé, mais qu'il ne reposait que sur une généralisation des observations humaines concernant les comportements des machines. Une induction empirique qui serait formulée de la sorte « (A) jusqu'à aujourd'hui les machines n'ont généralement pas mis en œuvre X, (B) donc les machines ne sont pas capables de X », avec l'établissement de la conclusion (B) comme une règle générale induite à partir de l'observation (A). L'histoire a, par de nombreuses occasions, donné tort à ce type de raisonnement inductif. Pendant longtemps il était généralement admis qu'une machine ne serait pas capable de battre l'individu humain aux échecs, puis cela est arrivé. Une machine ne semblait pas capable de battre un humain aux jeux de Go, puis cela est arrivé. Une machine ne semblait pas capable de manipuler le langage naturel pour faire de l'ironie, puis cela est arrivé. Une machine ne semblait pas capable de composer une symphonie puis cela est arrivé. etc.

Enfin, pour terminer cette discussion de l'intelligence générale artificielle, nous saisissons l'occasion qui nous est donnée de nous exprimer sur le sujet pour énoncer les arguments en faveur de la recherche d'une IA de niveau humain.

La première raison que nous pourrions donner pour justifier l'intérêt de concevoir une intelligence artificielle de niveau humain est celle qui nous tient le plus à cœur, elle est épistémologique. Il semble raisonnable de penser que concevoir un tel système permettra d'avancer vers une meilleure compréhension de l'intelligence humaine. La démarche est alors une opportunité de fédérer plusieurs disciplines scientifiques autour d'un objectif commun, un objectif fondamental des sciences humaines: celui d'essayer de comprendre, nos comportements, nos jugements, nos décisions et finalement nous-mêmes. En cela chercher à construire une HL-AI n'est pas qu'une opportunité technologique, mais aussi un vecteur par lequel il nous est donné d'étudier l'Homme et d'acquérir des connaissances essentielles à la bonne organisation de nos sociétés et des activités qui les composent.

Sur le plan technologique, la construction d'une intelligence artificielle de niveau humain n'est pour le moment qu'un horizon, un graal potentiellement inatteignable. Néanmoins, les chemins qui mènent à cet horizon sont pavés de découvertes de techniques et méthodes de traitement de l'information. Ces avancées, qu'elles nous rapprochent ou non de l'objectif, seront très probablement utiles en elle-même. Prises isolément, elles permettront et permettent déjà de résoudre d'autres problématiques applicatives contemporaines. Par exemple, l'entreprise DeepMind de Google est explicitement positionnée en tant qu'institution ayant l'objectif de concevoir des intelligences générales artificielles. Leurs travaux ont permis de développer des systèmes responsables d'avancées considérables en IA comme AlphaGo, premier système capable de battre tous les humains au jeu de Go, ou encore MuZero, un programme capable de jouer parfaitement à une grande variété de jeux vidéo sans en connaître les règles et sans avoir recours à des données humaines. Que ces programmes soient des intelligences générales artificielles est discutable, qu'ils fassent avancer la machine vers l'acquisition d'une intelligence équivalente à celle de l'Homme également. D'une certaine manière, les conclusions de ces débats importent peu. Les technologies découvertes pour la conception de ces programmes ont déjà eu d'autres utilités bien plus concrètes. AlphaFold, une version modifiée de AlphaGo, a révolutionné l'activité de prédiction de la structure des protéines et ouvre de nombreuses portes à l'industrie pharmacologique [Jumper et al. 2021]. MuZero a permis d'améliorer les méthodes de compression vidéo alors que le domaine ne faisait plus trop d'avancées depuis quelques

années [DeepMind 2022]. D'autres programmes basés sur ces technologies permettent d'améliorer les diagnostics médicaux [DeepMind 2016] ou encore d'optimiser la consommation énergétique des systèmes de refroidissement des data centers, réduisant ainsi leur impact écologique [DeepMind 2018]. Des entreprises comme DeepMind et bien d'autres institutions académiques et industrielles, qui ont pour horizon de modéliser et de construire des systèmes artificiels dont l'intelligence se rapproche de celle de l'Homme, construisent des savoirs et industrialisent des savoir-faire qui permettent de construire de nombreux autres systèmes utiles à l'industrie et aux sociétés.

Conclusion

Au commencement de notre travail de recherche, nous nous étions fixé l'objectif d'établir un modèle de la cognition humaine. Nous escomptions de ce modèle qu'il puisse servir d'architecture de référence pour appréhender le fonctionnement des systèmes dotés d'intelligence artificielle d'aujourd'hui et de demain, puisque ces systèmes ont vocation à automatiser et à simuler des facultés cognitives humaines. Nous voulions également que ce modèle puisse servir de cadre conceptuel pour analyser le comportement des utilisateurs en intégrant des notions jugées complexes telles que les émotions, l'expérience subjective ou l'inconscient.

Dans cette optique, nous avons mené une étude transdisciplinaire en explorant tour à tour des disciplines comme la psychologie cognitive, la philosophie de l'esprit, la neurobiologie et l'intelligence artificielle. Nous nous sommes particulièrement attardés sur une théorie psychologique nommée Dual-Process Theory. La Dual-Process Theory propose de décomposer l'activité cognitive de l'individu sous deux aspects : l'un étant implicite, automatique et affecté par l'expérience et les émotions (système 1), et le second étant explicite, délibéré et relatif aux raisonnements logiques (système 2). En prenant cette théorie comme point de départ et en nous inspirant du formalisme des systèmes de traitement de l'information et de la cybernétique, nous avons alors conçu une architecture cognitive nommée IPSEL.

En 2018, notre démarche s'inscrivait dans une tendance nouvelle concernant la conception de système, celle consistant à combiner les approches connexionnistes et les approches symboliques pour concevoir des systèmes qualifiés d'hybrides. Cette tendance recevait alors une attention grandissante de la part des communautés de chercheurs en intelligence artificielle, notamment grâce à la popularisation du livre de Daniel Kahneman [Kahneman 2011].

En 2019 nous présentions notre positionnement à travers un article publié dans la revue française Intellectica [Fruchart & LeBlanc 2019]. En 2020, Daniel Kahneman, lauréat du prix Nobel, fut invité à la 34ème conférence AAAI pour discuter de la cognition des systèmes artificiels de demain avec Yann LeCun, Geoffrey Hinton et Yoshua Bengio, les derniers lauréats du prix Turing.

Tout au long de notre travail, le paradigme Système 1 / Système 2 a gagné en popularité auprès des communautés de recherche en IA, aboutissant à des propositions plus ou moins détaillées d'architectures telles que celle proposée par une équipe de recherche de IBM [Ganapini et al. 2022] et celle proposée par Yann LeCun [LeCun 2022] directeur de la recherche en IA du groupe Meta.

La proposition de LeCun est très proche de la nôtre, cependant notre modèle ayant fait l'objet d'une communication en 2020 à l'occasion de la 13ème conférence internationale en Intelligence générale artificielle [Fruchart & LeBlanc 2020], la proposition de LeCun n'est pas documentée dans notre état de l'art puisqu'elle n'a pas orienté notre conception. Nous fournissons néanmoins une brève analyse comparative en annexe.

La modélisation IPSEL que nous avons décrite dans ce document est une architecture cognitive qui intègre la Dual-Process Theory mais également d'autres théories des sciences cognitives. Nous avons cherché à décrire plus précisément la relation particulière entre le système 1 et le système 2 en nous inspirant des idées d'auteurs classiques tels que Baruch Spinoza, Arthur Schopenhauer et Carl Jung.

A notre connaissance, dans la limite de ce qui nous a été donné de voir, le paradigme actuel en intelligence artificielle semble placer le système 2, la raison, comme élément de référence de l'individu. A l'inverse, dans notre proposition nous plaçons le système 1 comme élément central de l'activité cognitive et considérons le système 2 comme un outil qui n'est pas toujours mobilisé. L'essentiel de notre travail a alors été de théoriser la manière dont le système 2 se construit à partir de l'activité du système 1 pour lui servir de système d'aide à la décision, et comment ces deux systèmes coopèrent pour mettre en œuvre une intelligence générale.

Nous avons ensuite utilisé le modèle IPSEL pour établir une catégorisation à gros grain des systèmes artificiels. Nous proposons une classification en 5 catégories de systèmes qui se différencient par les fonctions cognitives qu'ils simulent. Toujours en utilisant le modèle IPSEL comme cadre de référence, nous proposons également 3 catégories de disposition cognitive dans lesquelles peut se trouver un individu lors de l'utilisation d'un système artificiel. Ces deux ensembles nous ont ensuite permis de différencier plusieurs situations d'interactions entre l'Homme et un système doté d'IA.

Nous avons également utilisé le paradigme que dépeint le modèle IPSEL pour enrichir les discussions concernant l'intelligence artificielle de confiance et l'intelligence générale artificielle.

Concernant l'intelligence artificielle de confiance, en 2018 la communauté semblait majoritairement avancer dans la direction de l'explicabilité. Depuis, plusieurs voix ont remis en question cette direction. Selon nous, deux questions résument bien cette prise de recul récente sur la pertinence des explications pour la problématique de la confiance Homme/IA : si l'explication permet d'avoir confiance dans le système, qu'est ce qui donne confiance dans l'explication ? Comment produire une explication compréhensible et suffisante pour l'utilisateur ? De notre opinion, le champ d'étude de l'explainable AI (XAI) déplace le problème mais ne le résout pas. Notre analyse argumente que la problématique de la confiance envers les systèmes dotés IA doit être abordée de manière globale. La solution ne peut pas être uniquement technique. Il est tout aussi important de considérer les aspects systémiques pour assurer une conception de confiance, et les facteurs humains pour permettre une utilisation en confiance.

Concernant l'intelligence générale artificielle, nous avons argumenté la possibilité de considérer une machine comme créative, consciente, sujette à des émotions, dotée d'un sens commun et d'une capacité de compréhension. Bien entendu ces arguments ne sont recevables que dans le spectre des définitions que nous avons données pour ces concepts. Une prérogative serait d'abord d'arriver à un consensus sur la nature de ces facultés chez l'individu humain. En cela, le paradigme que nous proposons peut servir à enrichir les discussions qui relatent ces capacités.

Bien que le modèle IPSEL puisse être reçu comme une spécification fonctionnelle de l'architecture cognitive d'une intelligence générale artificielle, il n'a pas vocation à être implémenté informatiquement. Nous proposons un pseudo-code du squelette algorithmique qu'il nous semblerait intéressant de tester ainsi que des directions technologiques qui pourraient permettre d'atteindre les traitements d'information que le modèle décrit. Néanmoins il existe encore plusieurs processus qui doivent être définis et différentes implémentations et paramétrages qui devraient être testés pour concevoir un système opérationnel.

Au-delà des promesses d'intelligence générale artificielle, la discipline de l'architecture cognitive permet de cristalliser des idées sur le fonctionnement de la cognition naturelle

dans le formalisme des systèmes de traitement de l'information. La discipline propose de définir des concepts et des principes dynamiques qui seront au mieux les plans des systèmes de demain, au pire des points d'ancrage autour desquels la communauté scientifique peut discuter. A travers notre proposition, notre volonté a principalement été d'essayer d'orienter les discussions sur l'IA de confiance et l'interaction Homme-IA vers un paradigme qui donne plus d'importance aux conditions humaines particulières : nos émotions, nos processus inconscients et la subjectivité de nos raisonnements.

Notre étude s'inscrit dans le cadre des réflexions sur le rapport entre l'Homme et l'outil. La confiance envers l'outil s'établira au fur et à mesure des usages et des bénéfices qu'ils engendrent. Néanmoins, pour garder le contrôle des systèmes dotés d'intelligence artificielle il sera nécessaire de ne pas s'y aliéner. L'utilisateur doit garder un esprit critique sur les performances et les limitations de son outil.

Perspectives

De notre point de vue, toutes les technologies sont là, le problème est désormais de trouver l'orchestration de ces techniques, l'intelligence générale artificielle pourrait n'être qu'un problème architectural. Pour implémenter l'architecture il faudra trouver les moyens physiques et matériels capables de mettre en œuvre ses principes (la réponse n'est pas nécessairement informatique mais pourrait nécessiter le recours à l'informatique quantique, à la biologie de synthèse ou à d'autres procédés). Il faudra également trouver l'environnement adéquat et le temps nécessaire pour entraîner le système, voire plusieurs générations de systèmes (il est potentiellement possible que seul l'environnement réel puisse permettre à un système artificiel de développer une intelligence générale).

Pour approfondir cette direction de recherche, si nous disposions des ressources nécessaires, nous envisagerions de mettre en place divers protocoles expérimentaux en facteurs humains pour tester la pertinence du modèle. Il faudrait en premier lieu tester la véracité des catégories de dispositions cognitives que nous avons énoncées. Il serait ensuite intéressant de trouver des moyens de mesurer l'état cognitif de l'utilisateur lors de l'utilisation pour permettre à la machine d'adapter les informations qu'elle envoie. Des mesures physiologiques ou l'analyse des micromouvements du corps pourraient fournir des indices clés sur lesquels le système se baserait pour savoir s'il a besoin d'expliquer ses résultats, de sortir l'utilisateur d'un effet tunnel ou de réduire ses communications pour soulager la charge cognitive de son opérateur.

Glossaire

AGI : Artificial general intelligence (intelligence générale artificielle)

ANI : Artificial narrow intelligence (intelligence artificielle restreinte)

DRL : Deep reinforcement learning (apprentissage par renforcement profond)

HL-AI : Human level artificial intelligence (intelligence artificielle de niveau humain)

IA : Intelligence artificielle

IPSEL : Information processing system with emerging logics

OS : Organes senseurs de l'architecture IPSEL

OE : Organes effecteurs de l'architecture IPSEL

S0 : Système Direct de l'architecture IPSEL

S1 : Système Intuitif de l'architecture IPSEL

S2 : Système Délibératif de l'architecture IPSEL

Bibliographie

[Adkins et al. 2022] Adkins, D., Alsallakh, B., Cheema, A., Kokhlikyan, N., McReynolds, E., Mishra, P., ... & Zvyagina, P. (2022, May). Method cards for prescriptive machine-learning transparency. In 2022 IEEE/ACM 1st International Conference on AI Engineering–Software Engineering for AI (CAIN) (pp. 90-100). IEEE.

[Aleksander & Burnett 1984] Aleksander, I., & Burnett, P. (1984). Reinventing man: The robot becomes reality. Holt, Rinehart & Winston.

[Anderson et al. 2004] Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. Psychological review, 111(4), 1036.

[Arnold 1960] Arnold, M. B. (1960). Emotion and personality.

[Arulkumaran 2019] Arulkumaran, K., Cully, A., & Togelius, J. (2019, July). Alphastar: An evolutionary computation perspective. In Proceedings of the genetic and evolutionary computation conference companion (pp. 314-315).

[Awad et al. 2020] Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J. F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. Proceedings of the National Academy of Sciences, 117(5), 2332-2337.

[Baars 1993] Baars, B. J. (1993). A cognitive theory of consciousness. Cambridge University Press.

[Baars 2005] Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Progress in brain research, 150, 45-53.

[Barber 1983] Barber, B. (1983). The logic and limits of trust.

[Bargh 1994] Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), Handbook of social cognition (pp. 1–40). Hillsdale, NJ : Erlbaum.

[Bengio 2017] Bengio, Y. (2017). The consciousness prior. arXiv preprint arXiv:1709.08568.

[Bengio 2022] Deleu, T., Góis, A., Emezue, C., Rankawat, M., Lacoste-Julien, S., Bauer, S., & Bengio, Y. (2022). Bayesian Structure Learning with Generative Flow Networks. arXiv preprint arXiv:2202.13903

[Berliner 1978] Berliner, H. J. (1978). A chronology of computer chess and its literature. *Artificial Intelligence*, *10*(2), 201-214.

[Binet & Simon 1916] Binet, A., & Simon, T. (1916). New methods for the diagnosis of the intellectual level of subnormals.(L'Année Psych., 1905, pp. 191-244).

[Booch et al. 2020] Booch, G., Fabiano, F., Horesh, L., Kate, K., Lenchner, J., Linck, N., ... & Srivastava, B. (2020). Thinking fast and slow in ai. arXiv preprint arXiv:2010.06002.

[Bostrom 2017] Bostrom, N. (2017). Superintelligence. Dunod.

[Brewer 1988] Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer (Eds.), Advances in social cognition (Vol. 1 , pp. 1–36). Hillsdale, NJ : Erlbaum.

[Brizendine 2006] Brizendine, L. (2006). The female brain. Broadway Books.

[Brundage et al. 2018] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Anderson, H. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

[Cannon 1920] Cannon, W. B. (1920). A laboratory course in physiology. Harvard University Press.

[Carnap 1928] Carnap, R. (1928). The Logical Structure of the World. Berkeley, University of California Press, 1967.

[Cellan-Jones 2014] Article de presse rédigé par Cellan-Jones, R. (2014).

[Chaiken 1987] Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Zanna, J. M. Olson, & C. P. Herman (Eds.), Social influence: The Ontario Symposium (Vol. 5 , pp. 3–39). Hillsdale, NJ : Erlbaum

[Chen et al. 2020] Chen, Z., Bei, Y., & Rudin, C. (2020). Concept whitening for interpretable image recognition. Nature Machine Intelligence, 2(12), 772-782.

[Chen et al. 2021] Chen, X., Yao, L., McAuley, J., Zhou, G., & Wang, X. (2021). A survey of deep reinforcement learning in recommender systems: A systematic review and future directions. arXiv preprint arXiv:2109.03540

[Chollet 2019] Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547.

[Ciravegna et al. 2021] Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Lió, P., Maggini, M., & Melacci, S. (2021). Logic explained networks. arXiv preprint arXiv:2108.05149.

[Claverie & Desclaux 2015] Claverie, B., & Desclaux, G. (2015). Commande, contrôle, communication: gestion cybernétique de systèmes d'information. *Hermès, La Revue*, (1), 70-77.

[Claverie & Desclaux 2016] Claverie, B., & Desclaux, G. (2016). C2-command and control: un système de systèmes pour accompagner la complexité. *Communication et organisation. Revue scientifique francophone en Communication organisationnelle*, (50), 255-278.

[Conrey et al. 2005] Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance . Journal of Personality and Social Psychology , 89 , 469–487.

[Crevier 1993] Crevier, D. (1993). AI: the tumultuous history of the search for artificial intelligence. Basic Books, Inc..

[Damasio 1996] Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 351(1346), 1413-1420.

[Damasio 2003] Damasio, A.R. (2003). Spinoza avait raison Joie et tristesse, le cerveau des émotions. Paris, Odile Jacob.

[Damasio 2006] Damasio, A. R. (2006). L'erreur de Descartes: la raison des émotions. Odile Jacob.

[Damasio 2017] Damasio, A. R. (2017). L'Ordre étrange des choses: La vie, les sentiments et la fabrique de la culture. Odile Jacob.

[DeepMind 2016] Article de blog rédigé par la rédaction de DeepMind
https://www.deepmind.com/blog/announcing-deepmind-health-research-partnership-with-moorfields-eye-hospital

[DeepMind 2018] Article de blog rédigé par la rédaction de DeepMind
https://www.deepmind.com/blog/safety-first-ai-for-autonomous-data-centre-cooling-and-industrial-control

[DeepMind 2022] Article de blog rédigé par la rédaction de DeepMind
https://www.deepmind.com/blog/muzeros-first-step-from-research-into-the-real-world

[Dehaene & Changeux 2005] Dehaene, S., & Changeux, J. P. (2005). Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. PLoS biology, 3(5), e141.

[Dehaene et al. 2017] Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it?. Science, 358(6362), 486-492.

[Deutsch 1973] Deutsch, M. (1973). The resolution of conflict: Constructive and destructive processes. Yale University Press.

[Desclaux 2021] Desclaux, G. (2021). Communication et confiance entre les humains et les machines intelligentes dotées de fonctions autonomes. *Hermes, La Revue*, *88*(2), 192-196.

[Desclaux 2022] Desclaux, G. (2022). Trust Between Humans and Intelligent Machines and Induced Cognitive Biases.

[Domingos 2015] Domingos, P. (2015). The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books.

[Ehsan et al 2021] Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I., Muller, M., & Riedl, M. O. (2021). The who in explainable ai: How ai background shapes perceptions of ai explanations. arXiv preprint arXiv:2107.13509.

[Endsley 1987] Endsley, M. R. (1987, September). The application of human factors to the development of expert systems for advanced cockpits. In Proceedings of the Human Factors Society Annual Meeting (Vol. 31, No. 12, pp. 1388-1392). Sage CA: Los Angeles, CA: SAGE Publications.

[Epstein 1973] Epstein, S. (1973). The self-concept revisited. Or a theory of a theory . American Psychologist , 28 , 404–416.

[Epstein 1994] Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious . American Psychologist , 49 , 709–724.

[Euler 1775] Euler, L. (1775). Lettres à une princesse d'Allemagne sur divers sujets de physique & de philosophie. Tome premier-[troisième]. Chez Steidel et Compagnie.

[Fazio 1990] Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework . Advances in Experimental Social Psychology , 23 , 75–109 .

[Fodor 1986] Fodor, J. A. (1986). La modularité de l'esprit, traduction Gerschenfeld. Paris: Ed. de Minuit.

[Freud 1905] Freud, S. (1905). Le trait d'esprit et sa relation à l'inconscient. Oeuvres complètes, 7.

[Freud et al. 1954] Freud, S., Bonaparte, M. E., Freud, A. E., Kris, E. E., Mosbacher, E. T., & Strachey, J. T. (1954). The origins of psycho-analysis: Letters to Wilhelm Fliess, drafts, and notes: 1887-1902.

[Fruchart & LeBlanc 2019] Fruchart, B., & Leblanc, B. (2019). Architecture cognitive et comportements. *Intellectica*, *71*(2), 139-156.

[Fruchart & LeBlanc 2020] Fruchart, B., & Blanc, B. L. (2020, September). Cognitive Machinery and Behaviours. In *International Conference on Artificial General Intelligence* (pp. 121-130). Springer, Cham.

[Ganapini et al. 2022] Ganapini, M. B., Campbell, M., Fabiano, F., Horesh, L., Lenchner, J., Loreggia, A., ... & Venable, B. (2022). Combining Fast and Slow Thinking for Human-like and Efficient Navigation in Constrained Environments. *arXiv preprint arXiv:2201.07050*.

[Garcez & Besold 2019] d'Avila Garcez, A. S.; and Besold, T. R. 2019. Special Issue: Neural-Symbolic Learning and Reasoning (NeSy'18). IfCoLog Journal of Logics and their Applications 6(4): 609– 610.

[Garcez & Lamb 2020] Garcez, A. D. A., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd Wave. arXiv preprint arXiv:2012.05876.

[Gawronski & Creighton 2013] Gawronski, B., & Creighton, L. A. (2013). Dual process theories.

[Haidt 2001] Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review, 108, 814– 834.

[Hamilton 2017] Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in neural information processing systems, 30.

[Hao 2020] K. Hao. We read the paper that forced Timnit Gebru out of Google. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

[Heaven 2020] W. D. Heaven, GPT-3 is shockingly good, and completely mindless. MIT Technology Review. https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/

[Hebb 2005] Hebb, D. O. (2005). The organization of behavior: A neuropsychological theory. Psychology Press.

[Hernandez 2017] Hernández-Orallo, J. (2017). Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artificial Intelligence Review, 48(3), 397-447.

[Higgins 2022] Higgins, I., Racanière, S., & Rezende, D. (2022). Symmetry-Based Representations for Artificial and Biological General Intelligence. Frontiers in Computational Neuroscience, 28.

[Hoefl et al. 2009] Hoefl ing, A., Likowski, K. U., Deutsch, R., H ä fner, M., Seibt, B., M ü hlberger, A., Weyers, P., & Strack, F. (2009). When hunger finds no fault with moldy corn: Food deprivation reduces food-related disgust . Emotion , 9 , 50–58.

[Hofmann et al. 2007] Hofmann, W., Rauch, W., & Gawronski, B. (2007). And deplete us not into temptation: Automatic attitudes, dietary restraint, and self-regulatory resources as determinants of eating behavior . Journal of Experimental Social Psychology , 43 , 497–504.

[Hopfield 1982] Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 79(8), 2554-2558.

[Jacoby 1991] Jacoby, L. L. (1991). A process-dissociation framework: Separating automatic from intentional uses of memory . Journal of Memory & Language , 30 , 513–541.

[Jaquet 2009] Jaquet, C. (2009). Le Spinoza protobiologiste de Damasio. In C. Jaquet, Pascal Sévérac & A.l Suhamy (éds.), La théorie spinoziste des rapports corps/ esprit et ses usages actuels. Paris, Hermann.

[James & Burkhardt 1975] James, W., & Burkhardt, F. (1975). Pragmatism (Vol. 1). Harvard University Press

[James 1884] James, W. (1884). What is an Emotion?(188-205). Mind, 9, 34.

[Ji et al. 2021] Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S. Y. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems.

[Jumper et al. 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

[Jung 1964] Jung, C.G. (1964). Man and His Symbols. Laurel.

[Kaelbling 1996] Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of artificial intelligence research, 4, 237-285.

[Kahneman 1979] Kahneman, D. (1979). Prospect theory: An analysis of decisions under risk. Econometrica, 47, 278.

[Kahneman 1994] Kahneman, D. (1994). New challenges to the rationality assumption. Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft, 18-36.

[Kahneman 2003] Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. American psychologist, 58(9), 697.

[Kahneman 2011] Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

[Kakas 1992] Kakas, A. C., Kowalski, R. A., & Toni, F. (1992). Abductive logic programming. Journal of logic and computation, 2(6), 719-770.

[Kempter et al. 1999] Kempter, R., Gerstner, W., & Van Hemmen, J. L. (1999). Hebbian learning and spiking neurons. Physical Review E, 59(4), 4498.

[Kinsey et al. 2003] Kinsey, A. C., Pomeroy, W. R., & Martin, C. E. (2003). Sexual behavior in the human male. American Journal of Public Health, 93(6), 894-898.

[Koening & Pearson 2019] Koenig-Robert, R. & Pearson, J. (2019). Decoding the contents and strength of imagery before volitional engagement. Scientific Reports, 9(1), 350

[Kotseruba & Tsotos 2016-1] Kotseruba, I., & Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Core cognitive abilities and practical applications. arXiv preprint arXiv:1610.08602.

[Kotseruba & Tsotos 2016-2] Kotseruba, I., Gonzalez, O. J. A., & Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Focus on perception, attention, learning and applications. arXiv preprint arXiv:1610.08602, 1-74.

[Kramer 1999] Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. Annual review of psychology, 50(1), 569-598.

[Kruglanski 2006] Kruglanski, A. W., & Dechesne, M. (2006). Are associative and propositional processes qualitatively distinct? Comment on Gawronski and Bodenhausen (2006) . Psychological Bulletin , 132 , 736–739

[Lacan 2013] Lacan, J. (2013). The ethics of psychoanalysis 1959-1960: The seminar of Jacques Lacan. Routledge.

[Laird et al. 2017] Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. Ai Magazine, 38(4), 13-26.

[Lecue 2019] Lecue, F. (2019). On the role of knowledge graphs in explainable AI. Semantic Web, (Preprint), 1-11

[LeCun & Bengio 1995] LeCun, Y. & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In M.A. Arbib (éd.) The Handbook of Brain Theory and Neural Networks. Cambridge, Mass., MIT Press.

[LeCun 2022] LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence Version 0.9. 2, 2022-06-27.

[Lee et al. 2004] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.

[Lee & Yoon 2021] Lee, D., & Yoon, S. N. (2021). Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *International Journal of Environmental Research and Public Health*, *18*(1), 271.

[Legg & Hutter 2007] Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. Frontiers in Artificial Intelligence and applications, 157, 17.

[Legg & Hutter 2007-2] Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. Minds and machines, 17(4), 391-444.

[Lemos 2020] Lemos, H., Avelar, P., Prates, M., Garcez, A., & Lamb, L. (2020, September). Neural-Symbolic Relational Reasoning on Graph Models: Effective Link Inference and Computation from Knowledge Bases. In International Conference on Artificial Neural Networks (pp. 647-659). Springer, Cham.

[Libet 1985] Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. Behavioral and Brain Sciences, 8(4), 529-539.

[Lindsay & Jacoby 1994] Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process-dissociations: The relationship between facilitation and interference . Journal of Experimental Psychology: Human Perception & Performance , 20 , 219–234.

[Liu 2022] Liu, S., Marris, L., Hennes, D., Merel, J., Heess, N., & Graepel, T. (2022). NeuPL: Neural Population Learning. arXiv preprint arXiv:2202.07415.

[Macron 2018] Allocution du président de la République Française E. Macron (2018) : AI for Humanity. https://www.elysee.fr/emmanuel-macron/2018/03/29/discours-du-president-de-la-republique-sur-lintelligence-artificielle

[Maharaj et al. 2022] Maharaj, S., Polson, N., & Turk, A. (2022). Chess AI: competing paradigms for machine intelligence. *Entropy*, *24*(4), 550.

[Marcus 2020] Marcus, G. (2020). The next decade in ai: four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177.

[Mayer 1995] Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709-734.

[McCulloch & Pitts 1943] McCulloch, W.S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115-133.

[Menon 2011] Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. Trends in cognitive sciences, 15(10), 483-506

[Meyerson 2011] Meyerson, É. (2011). Du cheminement de la pensée. Vrin.

[Miller et al. 2017] Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. arXiv preprint arXiv:1712.00547.

[Miller 2019] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.

[Minsky 1986] Marvin Minsky. The Society of Mind. Simon & Shuster, Inc., New York, NY, 1986.

[Mnih 2015] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. nature, 518(7540), 529-533.

[Moir & Jessel 1991] Moir, A., & Jessel, D. (1991). Brain sex: The real difference between men and women (p. 256). London: Mandarin.

[Monier 2019] Monier, C. (2018). Les neurosciences au sein des sciences de la cognition–Une revue de la littérature plaidant pour un devenir spinoziste des neurosciences. Intellectica, 69(1), 27-132.

[Muir 1987] Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. International journal of man-machine studies, 27(5-6), 527-539.

[Myers 2002] Myers, D. G. (2002). Human Intuition: The Brain Behind the Scenes Intuition: Its Powers and Perils. Cerebrum, 4(3), 100-113.

[Neumann 1945] Von Neumann, J. [1945]. First Draft of a Report on the EDVAC. IEEE Annals of the History of Computing, 15(4), 27-75, 1993.

[Neumann 1948] Von Neumann, J. [1948]. Théorie générale et logique des automates : le mécanisme algorithmique de John von Neumann. Ceyzérieu, Éditions Champ Vallon, 1996.

[Newell & Simon 1959] Newell, A., Shaw, J.C. & Simon, H.A. (1959, June). Report on a general problem solving program. In IFIP Congress (Vol. 256, p. 64), IFIP.

[Newell 1980] Newell, A. (1980). Physical symbol systems. Cognitive science, 4(2), 135-183

[Newell 1994] Newell, A. (1994). Unified theories of cognition. Harvard University Press.

[Nietzsche 1968] Nietzsche, F. W. (1968). Basic writings of Nietzsche (Vol. 1). Random House Digital, Inc..

[Nori et al 2019] Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223.

[Otto 2021] Otto, F. (2021). Model-free deep reinforcement learning—algorithms and applications. In Reinforcement Learning Algorithms: Analysis and Applications (pp. 109-121). Springer, Cham.

[Pandl et al. 2020] Pandl, K. D., Thiebes, S., Schmidt-Kraepelin, M., & Sunyaev, A. (2020). On the convergence of artificial intelligence and distributed ledger technology: A scoping review and future research agenda. *IEEE access*, *8*, 57075-57095.

[Parasuraman 1997] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human factors, 39(2), 230-253.

[Pavlov 2010] Pavlov, P. I. (2010). Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. Annals of neurosciences, 17(3), 136.

[Payne 2008] Payne, B. K. (2008). What mistakes disclose: A process dissociation approach to automatic and controlled processes in social psychology . Social and Personality Psychology Compass , 2 , 1073–1092.

[Pearl 1998] Pearl, J. (1998). Graphical models for probabilistic and causal reasoning. Quantified representation of uncertainty and imprecision, 367-389.

[Piaget et al. 1925] Piaget, J., Rousseau, J., Piaget, M., Deslex, M., & Claparéde, E. (1925). Le langage et la pensée chez l'enfant.

[Pinker 2003] Pinker, S. (2003). The blank slate: The modern denial of human nature. Penguin.

[Rasmussen 1983] Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. IEEE transactions on systems, man, and cybernetics, (3), 257-266.

[Rempel et al. 1985] Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. Journal of personality and social psychology, 49(1), 95.

[Ribeiro et al. 2016] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[Ronen 2022] Ronen, M., Finder, S. E., & Freifeld, O. (2022). DeepDPM: Deep Clustering With an Unknown Number of Clusters. arXiv preprint arXiv:2203.14309.

[Rosenblatt 1958] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386.

[Rossi et al. 2021] Rossi, E., Kenlay, H., Gorinova, M. I., Chamberlain, B. P., Dong, X., & Bronstein, M. (2021). On the Unreasonable Effectiveness of Feature Propagation in Learning on Graphs with Missing Node Features. arXiv preprint arXiv:2111.12128.

[Rousseau 1998] Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. Academy of management review, 23(3), 393-404.

[Rudin 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215

[Rumelhart 1988] Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1988). Learning representations by back-propagating errors. Cognitive Modeling, 5(3), 1.

[Sacks 2018] Sacks, O. (2018). L'homme qui prenait sa femme pour un chapeau, et autres récits cliniques. Le Seuil.

[Schmidhuber 1991] Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In Proc. of the international conference on simulation of adaptive behavior: From animals to animats (pp. 222-227).

[Schmidhuber 2015] Schmidhuber, J. (2015). On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. arXiv preprint arXiv:1511.09249.

[Schopenhauer 1888] Schopenhauer, A. (1888). Le monde comme volonté et comme représentation (Vol. 61). Ancienne libr. Germer Baillière et Cie., Félix Alcan.

[Schrittwieser et al. 2020] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., ... & Silver, D. (2020). Mastering atari, go, chess and shogi by planning with a learned model. Nature, 588(7839), 604-609.

[Sharma et al. 2021] Sharma, A., Xu, K., Sardana, N., Gupta, A., Hausman, K., Levine, S., & Finn, C. (2021). Autonomous Reinforcement Learning: Formalism and Benchmarking. arXiv preprint arXiv:2112.09605.

[Sherman 2008] Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses . Psychological Review , 115 , 314–335.

[Silver et al. 2017] David, S., Julian, S., Karen, S., Ioannis, A., Aja, H., Arthur, G., ... & Yutian, C. (2017). Lillicrap Timothy P., Hui Fan, Sifre Laurent, van den Driessche George, Graepel Thore, Hassabis Demis. Mastering the game of Go without human knowledge, Nat, 550(7676), 354-359.

[Simon 1972] Simon, H. A. (1972). Theories of bounded rationality. Decision and organization, 1(1), 161-176.

[Smith & DeCoster 2000] Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems . Personality and Social Psychology Review , 4 , 108–131.

[Spinoza 1677] Spinoza, B. (1677). Traité de la Réforme de l'Entendement, trad. de A. Koyré. Paris, Vrin,1984.

[Srivastava et al. 2022] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Kim, H. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

[Stanovich & West 2000] Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. Behavioral and brain sciences, 23(5), 645-665.

[Steiner 2011] SteineR, P. (2011). Énaction, pragmatisme et minimalisme représentationnel. Peut-on se passer de représentations en sciences cognitives.

[Stewart et al. 2010] Stewart, J., Gapenne, O. & Di Paolo, E.A. (2010). Enaction: Toward a New Paradigm for Cognitive Science. Cambridge, MA, The MIT Press.

[Stocco et al. 2019] Stocco, A., Steine-Hanson, Z., Koh, N., Laird, J. E., Lebiere, C. J., & Rosenbloom, P. (2019). A Common Architecture for Human and Artificial Cognition Explains Brain Activity Across Domains. bioRxiv, 703777.

[Strack & Deutsh 2004] Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior . Personality and Social Psychology Review , 8 , 220–247.

[Strubell 2019] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

[Sun 2003] R. Sun, (2003). A Tutorial on CLARION. Technical report, Cognitive Science Department, Rensselaer Polytechnic Institute.

[Sun 2004] Sun, R. (2004). Desiderata for cognitive architectures. Philosophical Psychology, 17(3), 341-373.

[Sun 2007] Sun, R. (2007). The importance of cognitive architectures: An analysis based on CLARION. Journal of Experimental & Theoretical Artificial Intelligence, 19(2), 159-193.

[Sutton & Barto 2018] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

[Tang et al. 2022] Tang, Y., Tian, Y., & Ha, D. (2022). EvoJAX: Hardware-Accelerated Neuroevolution. arXiv preprint arXiv:2202.05008.

[Tazrout 2022] Article de presse rédigé par la rédaction de ActuIA, Tazrout, Z. (2022). https://www.actuia.com/actualite/le-marche-mondial-de-lintelligence-artificielle-en-entreprise-atteindra-342-milliards-de-dollars-en-2022/

[Tual 2019] Article de presse rédigé pour la rédaction du Monde, par Tual, M. (2019). https://www.lemonde.fr/pixels/article/2019/10/30/intelligence-artificielle-un-programme-se-hisse-parmi-les-meilleurs-joueurs-de-starcraft-ii_6017484_4408996.html

[Turing 1950] Turing, A.M (1950). Computing Machinery and Intelligence. Mind, 49, 433-460.

[Varela et al. 1993]Varela Francisco, J., Evan, T., & Eleanor, R. (1993). L'inscription corporelle de l'esprit. Sciences cognitives et expérience humaine.

[Villani et al. 2018] Villani, C., Bonnet, Y., Berthet, C., Levin, F., Schoenauer, M., Cornut, A. C., & Rondepierre, B. (2018). Donner un sens à l'intelligence artificielle: pour une stratégie nationale et européenne. Conseil national du numérique.

[Whitehead & Russell 1910] Whitehead, A. N., & Russell, B. (1910). Principia mathematica. Cam-bridge. Press, Cambridge (1910-1913).

[Whitehead 1962] Whitehead, A. N. (1962). Symbolism, its meaning, and effect.

[Wiegmann 2002] Wiegmann, D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. Human Factors, 44(1), 44-50.

[Wiener 1948] Wiener, N. [1948]. Cybernetics or Control and Communication in the Animal and the Machine. Paris, Hermann & Cie & Cambridge, Mass., MIT Press, 1965.

[Wikipedia 1] Définition de la cognition. https://fr.wikipedia.org/wiki/Cognition.

[Wu et al. 2020] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32(1), 4-24.

[Wundt 1904] Wundt, W. M. (1904). Principles of physiological psychology (Vol. 1). Sonnenschein.

[Zajonc 1998] Zajonc, R. B. (1998). Emotions.

[Zhang et al. 2022] Zhang, B., Dreksler, N., Anderljung, M., Kahn, L., Giattino, C., Dafoe, A., & Horowitz, M. C. (2022). Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers. *arXiv preprint arXiv:2206.04132*.

[Zhang & Yao 2022] Zhang, Y., & Yao, Q. (2022, April). Knowledge Graph Reasoning with Relational Digraph. In Proceedings of the ACM Web Conference 2022 (pp. 912-924).

[Zhu 2022] Zhu, X., Li, Z., Wang, X., Jiang, X., Sun, P., Wang, X., ... & Yuan, N. J. (2022). Multi-Modal Knowledge Graph Construction and Application: A Survey. arXiv preprint arXiv:2202.05786.

[Zimmerman 2018] Interview J. Zimmerman in « An Ingenious Approach to Designing AI That Doctors Trust » (2018). Fastcompany.com.
https://www.fastcompany.com/90157144/an-ingenious-approach-to-designing-ai-that-doctors-trust

[Zuboff 1988] Zuboff, S. (1988). In the age of the smart machine: The future of work and power. Basic Books, Inc..

Annexes 1 : Comparaison entre JEPA et IPSEL

Yann LeCun, directeur de la recherche en IA pour le groupe Meta et lauréat du prix Turing, a récemment proposé l'architecture cognitive d'un agent artificiel autonome ayant vocation à développer une intelligence générale artificielle que LeCun présente comme « capable d'apprendre et de raisonner comme les animaux et les humains ». [LeCun 2022]. Sa proposition est très semblable à la nôtre qui la précède de deux ans.

Le modèle de LeCun est nommé JEPA. JEPA et IPSEL sont deux propositions de l'architecture cognitive d'un agent qui apprend de manière autonome (self-supervised learning). Pour guider cet auto-apprentissage, les deux modèles contiennent un module qui effectue des évaluations des états de l'agent (Cost Module pour JEPA, Émotions pour IPSEL) et guide l'apprentissage.

Les deux modèles spécifient plusieurs niveaux de traitement de l'information. La distinction la plus importante est celle faite entre un niveau rapide réactif/automatique et un niveau plus lent qui s'apparente au raisonnement. Il s'agit là de l'intégration de la théorie psychologique du *Dual-Process* (Mode1/Mode2 pour JEPA, S1/S2 pour IPSEL). Cependant, IPSEL décrit également un troisième niveau, le système Direct (S0) isolé des deux autres, que le modèle JEPA ne décrit pas formellement mais que l'on pourrait retrouver exprimé dans un autre vocabulaire, cependant puisque LeCun n'en parle pas nous ne nous aventurons pas à le faire à sa place. Les modules qui forment les deux architectures sont relativement similaires : module de perception, module d'action, module d'évaluation, plusieurs mémoires correspondant aux différents modes de traitement. Cependant l'architecture de ces modules, les liens qu'ils partagent les uns avec les autres, ne sont pas exactement identiques entre JEPA et IPSEL (Cf figure ci-dessous).

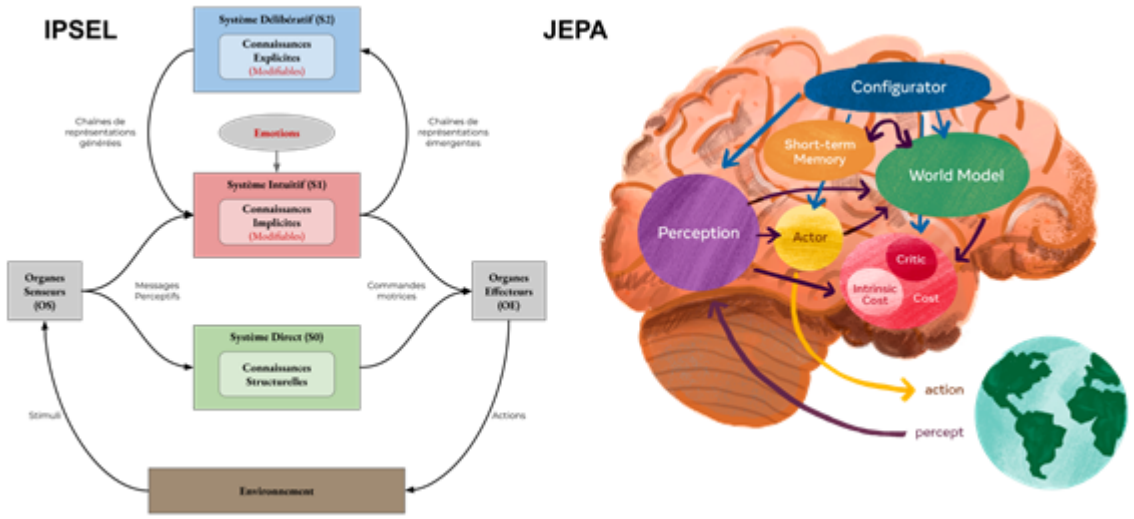

Figure annexe 1 : *Architecture IPSEL (Bryan Fruchart) et JEPA (Yann LeCun)*

Dans JEPA comme dans IPSEL, les comportements récurrents initié par le mode 2 de JEPA, respectivement le système 2 de IPSEL, deviennent des connaissances du mode 1 dans JEPA, respectivement du système 1 dans IPSEL, qui peut ensuite les mettre en œuvre sur un mode nommé réactif chez JEPA, respectivement intuitif chez IPSEL.

Dans JEPA, pour le traitement du mode 1, le chemin «réactif» qui lie les perceptions aux actions ne prend pas en considération les connaissances sur le monde ni le module d'évaluation. A l'inverse, dans IPSEL, les traitements automatiques du S1 prennent en considération les connaissances implicites du monde, les émotions (module d'évaluation) et le résultat des traitements du S2. Sur ce point, dans l'architecture JEPA les modes 1 et 2 semblent être indépendants, en phase de marche c'est l'un des deux modes qui est appelé et commande les actions. Dans le cas de IPSEL, les systèmes 1 et 2 sont interdépendants, le système 1 est le seul à commander les actions et prend les productions du S2 comme sources d'information supplémentaire. Dit autrement, il semblerait que les modes 1 et 2 sont en compétition dans JEPA, alors que dans IPSEL ils sont en collaboration.

Dans JEPA, le mode 2 produit des séquences d'actions en utilisant des évaluations de l'état simulé du monde. Dans IPSEL, le système 2 produit des simulations du monde par le biais de séquences de représentations explicites. A notre sens, le mode 2 de JEPA ne peut produire des discours introspectifs, ou accomplir des fonctions d'imagination ou de

souvenir car ses processus apprennent et sont optimisés pour produire des actions. L'Homme ne pense pas uniquement à des actions et nous ne comprenons pas comment l'architecture JEPA permet ces autres modes de pensées qui n'ont pas la mise en mouvement du corps comme finalité.

De manière générale, ce que IPSEL semble apporter de plus que JEPA est la véritable distinction entre les traitements des deux niveaux. Dans JEPA les deux niveaux semblent utiliser le même format d'information, puisque les deux niveaux peuvent contrôler l'unique module d'action et que les productions du mode 2 peuvent être directement assimilées par le mode 1 qui les mettra lui-même en œuvre plus tard. Chez JEPA la différence entre mode 1 et mode 2 semble tenir uniquement sur la quantité de calcul opéré. La volonté semble être de permettre à un mode d'opérer plus rapidement que l'autre. Chez IPSEL, entre 1 et 2 ce ne sont tout simplement pas les mêmes calculs et leur différence de vitesse n'est pas une finalité recherchée mais une conséquence. La finalité recherchée chez IPSEL est l'intégration de deux formats de connaissances, implicites et explicites, que nous ne retrouvons pas formellement dans JEPA.

Annexes 2 : Catégorisation des systèmes et des utilisateurs

Etape 1 : Identifier la catégorie générique du système artificiel (CS)

Catégorie	Caractéristiques	Éléments nécessaires à la compréhension des comportements du système	Analogie dans le modèle IPSEL / Nature des connaissances utilisées
0 - Classique	• Pas d'apprentissage	• Entrée • Traitement	S0 / Pas de connaissance
1 - Inductif	• Apprentissage statistique	• Entrée • Traitement • Entrainement	S1 / Connaissances implicites subjectives
2 - Déductif	• Apprentissage logique	• Entrée • Traitement • Logique d'inférence	S2 / Connaissances explicites Humaines
3 - Hybrid supervisé	• Apprentissage statistique et logique	• Entrée • Traitement • Entrainement • Logique d'inférence	S1&S2 / Connaissances implicites subjectives et connaissances explicites Humaines
4 - Hybrid émergent	• Apprentissage statistique et logique	• Entrée • Traitement • Entrainement • Logique d'inférence	S1&S2 / Connaissances implicites subjectives et connaissances explicites subjectives

Etape 2 : Identifier la disposition cognitive archétypique de l'utilisateur lors de l'utilisation du système (CU)

Disposition cognitive	Connaissances mobilisées	Traitement cognitifs dominant	Avantages	Inconvénients
E - Equilibré	• Implicites (savoir-faire) • Explicites (savoir)	Aucun	Polyvalence des comportements	Sujet aux doutes. Raisonnement biaisés par l'intuition. Intuition amoindrie par les raisonnement.
C - Contrôle intuitif	• Implicites (savoir-faire)	Système Intuitif (S1) / Intuition	Perception et contrôle moteur précis, rapide et fluide. Décision rapide.	Jugements et décisions non délibérés conceptuellement. Sourd aux arguments et explications rationnels
D - Décision Délibéré	• Explicites (savoir)	Système Délibératif (S2) / Raisonnement	Jugements et décisions délibérés conceptuellement. Disposé à recevoir et à intégrer arguments et explications rationnels	Perception et contrôle moteur imprécis, lent et saccadés. Décision lente.

Étape 3 : Utiliser la catégorie du système artificiel et la disposition cognitive de l'utilisateur pour identifier la situation d'interaction et les variables (instruction, entraînement, contextualisation, explication) qui nécessitent plus d'attention (case orange dans le tableau ci-dessous).

	0 - Classique	1 - Inductif	2 - Déductif	3 - H.Supervisé	4 - H. Émergent
Pré-utilisation	Instruction	Instruction	Instruction	Instruction	Instruction
	Entrainement	Entrainement	Entrainement	Entrainement	Entrainement
Utilisation en mode Équilibrée	Contextualisation	Contextualisation	Contextualisation	Contextualisation	Contextualisation
	Explication	Explication	Explication	Explication	Explication
Pré-utilisation	Instruction	Instruction	Instruction	Instruction	Instruction
	Entrainement	Entrainement	Entrainement	Entrainement	Entrainement
Utilisation en mode Contrôle Intuitif	Contextualisation	Contextualisation	Contextualisation	Contextualisation	Contextualisation
	Explication	Explication	Explication	Explication	Explication
Pré-utilisation	Instruction	Instruction	Instruction	Instruction	Instruction
	Entrainement	Entrainement	Entrainement	Entrainement	Entrainement
Utilisation en mode Décision délibérée	Contextualisation	Contextualisation	Contextualisation	Contextualisation	Contextualisation
	Explication	Explication	Explication	Explication	Explication

Instruction : temps pré-utilisation pendant lequel l'utilisateur reçoit un ensemble de connaissances générales portant sur les fonctions du système, son fonctionnement, ses caractéristiques et des instructions concernant son utilisation. Durant ce processus, l'utilisateur reçoit essentiellement des connaissances explicites, c'est-à-dire un savoir qu'il lui sera permis de mobiliser lors de l'utilisation de système de ce type.

L'entraînement : temps pré-utilisation pendant lequel l'utilisateur construit un ensemble de connaissances implicites, un savoir-faire particulier. Durant l'entraînement l'utilisateur s'entraîne à utiliser un système particulier dans des conditions particulières.

Contextualisation : correspond à une action menée par l'utilisateur qui contextualise les productions du système pour évaluer leur pertinence. Il s'agit pour l'humain de considérer la fonction, le fonctionnement et le domaine d'entraînement du système pour relativiser ses productions. Les informations permettant la contextualisation peuvent être fournies par le système, un système tiers ou un collaborateur.

L'explication : correspond à une communication du système qui fournit des éléments explicatifs à l'utilisateur. Intégrant ces éléments, l'utilisateur doit pouvoir comprendre les traitements qui ont amené le système à produire son résultat. Il s'agit pour le système de

présenter un raisonnement causal explicitant les causes qui l'ont mené à générer les productions comme effets. L'explication peut être fournie par le système, un système tiers ou un collaborateur.

Étape 4 : Schéma décomposant l'état de confiance en deux catégories, la confiance délibérée et la confiance intuitive, et identification des variables affectant leurs états.