

Synthèse de la présentation du problème du sens commun en IA

1) Le bon sens : la matière noire de l'intelligence

Aujourd'hui l'intelligence artificielle fascine par sa capacité à manipuler le langage avec fluidité, mais elle bute paradoxalement sur des concepts physiques et sociaux qu'un enfant en bas âge maîtrise intuitivement, qu'on appelle plus généralement le **bon sens**. Pour nous situer, un peu de contexte. Bien avant que l'intelligence artificielle ne s'empare de cette notion, Descartes en donnait déjà une définition célèbre, en 1637, dans le *Discours de la méthode* : « Le bon sens est la chose du monde la mieux partagée [...]. Car ce n'est pas assez d'avoir l'esprit bon, mais le principal est de l'appliquer bien. »

Autrement dit, chacun de nous en a reçu, mais encore faut-il savoir s'en servir. À travers cette définition, on peut déjà saisir ce qui caractérise le bon sens : quelque chose que nous possédons tous, et que nous utilisons sans même y penser.

Le bon sens est la matière noire de la pensée. C'est cet immense volume de connaissances sur le monde (la gravité, la solidité, les réactions sociales basiques) que nous possédons tous, mais que nous ne disons ou n'écrivons jamais car c'est trop évident. Le problème des IA actuelles vient de là : elles ingèrent toutes les données qui se trouvent sur Internet pour essayer d'imiter la cognition humaine, or les humains n'écrivent jamais sur l'ordinaire mais parlent davantage de l'extraordinaire et donc l'évident n'est écrit nulle part. Les machines ont donc d'énormes trous dans leur compréhension fondamentale de la réalité.

2) De l'échec symbolique à l'illusion statistique

Pour combler ce vide, la recherche en IA s'est d'abord épuisée à vouloir coder manuellement chaque règle du monde avec notamment Lenat en 1984 et son benchmark **Cyc**. Malheureusement, écrire toutes ces connaissances à la main demande énormément de travail, et tout ne peut pas être couvert.

Les modèles actuels semblent avoir surmonté les échecs historiques avec brio, en atteignant des scores historiques de plus de 95%, dépassant même la référence humaine, en générant des réponses politiquement correctes, fluides et rapides. Cependant, les chercheuses et chercheurs ont démontré que cette performance relève davantage d'une imitation **statistique** fragile que d'une véritable **compréhension** du monde physique. La question de comment évaluer et quantifier le bon sens se pose d'autant plus quand les **benchmarks** sont finalement réduits à des bases de données apprises par cœur. Leur capacité d'adaptation reste extrêmement vulnérable à la moindre reformulation, inversion de phrase, ou ajout de bruit (détails), prouvant ainsi qu'elles manipulent des statistiques textuelles plutôt que des concepts. Les chercheurs ont également constaté que leur accorder plus de temps de réflexion pour se confronter à leurs propres contradictions et se corriger seuls, n'améliore que le calcul et non l'intuition physique fondamentale qui leur fait défaut.

3) Les voies d'avenir

Le bon sens en IA est un thème toujours d'actualité et le monde de la recherche cherche encore la solution la plus adaptée. Pour sortir de l'impasse purement statistique, la communauté scientifique n'explore plus une seule solution miracle, mais s'oriente plutôt vers plusieurs stratégies complémentaires. D'un côté certains retournent avec Brooks et affirment que c'est l'**IA incarnée** qui va régler ce problème de bon sens en affirmant qu'une machine a besoin de vivre certaines choses pour les comprendre. En parallèle les **systèmes multi-agents** devraient aider à pallier la manque de normes sociales des machines en faisant émerger naturellement le bon sens des interactions entre agents.

D'un autre côté, on peut penser à l'**IA neuro-symbolique** comme le préconise la chercheuse Choi Yejin avec son projet COMET. Elle préconise une approche neuro-symbolique basée sur le langage naturel pour doter les machines d'un raisonnement génératif et flexible. Enfin, une piste tout aussi prometteuse serait l'**IA développementale** ou "la machine enfant". L'idée est de doter la machine d'une "motivation intrinsèque" (la curiosité). L'agent génère ses propres buts et explore son environnement, comme un enfant.

4) Au cœur des échanges

Les échanges avec le public se sont d'abord articulés autour de la nature et de l'origine du bon sens, défini comme un ensemble de règles **implicites** et non orales. Les participants se sont interrogés sur le fait que le bon sens nous est transmis non pas par notre propre **expérience** mais par l'expérience de toutes les sociétés avant nous. La question s'est donc posée : faut-il attendre plusieurs siècles pour que des sociétés d'IA se forgent ce bon sens elles aussi ?

Le parallèle avec le développement cognitif a permis de souligner que, chez l'humain et même le bébé, la capacité de généralisation semble innée et intimement liée à un instinct de survie.

Face à ce constat, le public a questionné le fonctionnement de la machine, se demandant si la quête du bon sens artificiel ne se réduisait pas, au fond, à un simple **apprentissage par renforcement**. Enfin, le débat s'est conclu sur une remise en question de l'objectif lui-même : le besoin de doter l'IA de bon sens est-il réellement pertinent ? Pour les participants, la conclusion est apparue contre-intuitive, arguant que le bon sens ne constitue finalement pas un critère déterminant pour accorder sa confiance à une intelligence artificielle.

En résumé, si le public a eu tendance à limiter le bon sens aux seules règles physiques de notre monde, ce concept est en réalité beaucoup plus vaste. C'est d'ailleurs le défi majeur des équipes de recherche aujourd'hui : comment transmettre une si grande quantité de normes implicites à une machine ?

Comment l'amener à comprendre cette infinité de situations sans avoir à toutes les lui programmer manuellement ?

5) Conclusion

De nos jours, l'intelligence artificielle est un prodige de syntaxe mais reste inexpérimentée quant à la sémantique du réel. L'illusion que les machines nous donnent grâce au par cœur et à leur grande capacité de calcul atteint ses limites. Pour les modèles de demain, la recherche ne vise plus une simple performance textuelle, mais une véritable robustesse face aux problématiques humaines. Les scientifiques l'ont en effet bien compris : on ne peut pas apprendre à une machine ce qui n'est écrit nulle part. Imiter le bon sens ne semblant pas fonctionner, les solutions de demain résident davantage dans le fait de comprendre et d'expérimenter le monde réel, car le bon sens ne peut pas se télécharger.

6) Bibliographie

- McCarthy, J. (1958/1959). *Programs with Common Sense* (L'Advice Taker). (Première tentative d'intégrer le bon sens via la logique formelle).
- Lenat, D. B. (1984). *Projet Cyc*. (Tentative d'encodage symbolique manuel des règles du monde).
- Brooks, R. (1990). *Elephants Don't Play Chess*. (Théorie fondatrice de l'IA incarnée : la machine doit agir physiquement dans le monde pour le comprendre).
- Kahneman, D. (2011). *Système 1 / Système 2 : Les deux vitesses de la pensée*. (Le bon sens vu comme une pensée rapide, instinctive, mais source potentielle de biais).
- Zellers, R., et al. (2019). *HellaSwag*. (Benchmark textuel conçu pour tromper les machines, aujourd'hui "résolu" par cœur par les grands modèles de langage).
- Davis, E. (2023). *Benchmarks for Automated Commonsense Reasoning: A Survey*. (Recensement de 139 tests textuels, visuels et physiques).
- Park, J. S., et al. (2023). *Generative Agents (Emergence World)*. (Simulations de sociétés d'agents IA pour observer l'émergence de normes sociales et de l'organisation temporelle).
- Choi, Y. (2019). *Projet COMET*. (Approche de l'IA neuro-symbolique liant la flexibilité du langage naturel aux graphes de connaissances).
- Bian, N., et al. (2023). *Études sur les limites des LLM face aux données temporelles*. (Études sur les limites des LLM, comme ChatGPT, face aux données liées au temps et aux interactions sociales).
- LeCun, Y. (2022). *La "matière noire de l'intelligence"*. (Métaphore pour désigner le socle de connaissances implicites invisibles dans les textes).

Fiche de synthèse : atelier « IA générative et hypertrucages »

Walid Benbouda, Mahmoud Dahir, Michel Junior Delanoue

M2 Intelligence Artificielle, UE IA & Cognition (INF2346M), Université Claude Bernard Lyon 1

Atelier présenté le 14 octobre 2026

Résumé. Cet atelier s'appuie sur le livre blanc *IA générative et hypertrucages* [1]. Nous montrons comment ces faux sont produits, pourquoi ils posent un problème qui dépasse la technique, et quelles réponses existent en 2026. Notre thèse : face à des contenus que nos sens ne savent plus authentifier, la confiance doit passer de *ce que l'on voit* à *d'où cela vient*.

Mots-clés : IA générative, hypertrucage (deepfake), désinformation, biais cognitifs, provenance, AI Act.

1. Contexte et problématique

Manipuler une image n'est pas nouveau : les retouches de photographies officielles existaient déjà sous Staline. Ce qui change avec l'IA générative, c'est l'échelle. Le livre blanc parle d'une « démocratisation accélérée » : depuis fin 2022, des millions de personnes sans compétence technique peuvent produire texte, image, voix ou vidéo réalistes [1]. Nous avons donc posé la question suivante : **quand voir ne suffit plus pour croire, sur quoi fonder la confiance dans un contenu numérique ?**

2. Comprendre : de la prédiction statistique à l'industrialisation du faux

Un modèle de langage (LLM) ne raisonne pas : il prédit le mot suivant le plus probable à partir de milliards de textes. C'est cette fluidité qui donne l'illusion d'une intelligence. Pour l'image, deux familles de modèles dominent. Les réseaux antagonistes génératifs (GAN, 2014) font s'affronter un faussaire et un expert jusqu'à ce que l'expert ne distingue plus le faux [12]. Les modèles de diffusion (Midjourney, DALL-E) partent d'un bruit aléatoire qu'ils « nettoient » étape par étape. Pour la voix, VALL-E imite une personne à partir de trois secondes d'audio [13]. Le livre blanc définit l'hypertrucage comme un média fabriqué ou manipulé pour imiter l'identité d'une personne. Il rappelle qu'il n'est pas malveillant par nature (cinéma, doublage, art). Le basculement tient en trois points : la qualité trompe désormais les experts, la création prend quelques secondes sur un téléphone, et la diffusion atteint des millions de personnes avant le premier démenti. La frise présentée (faux Obama en 2018, faux Zelensky en 2022, fraude de 25 M\$ par visioconférence truquée en 2024) montre ce passage de l'artisanat du faux à une industrie.

3. Le problème : des dégâts réels et un cerveau mal armé

Le livre blanc classe les usages en cinq familles (assistants, influence, expertise, création, doubles numériques). Dans chacune, le même outil sert à améliorer et à tromper [1]. Trois cas illustrent les dégâts. Au Gabon (2018-2019), une vidéo *authentique* du président Ali Bongo, crue truquée, a servi de prétexte à une tentative de coup d'État. Pour le livre blanc, ce n'est pas la manipulation qui crée le danger, mais « le climat de défiance établi » [1, p. 22]. En 2023, une fausse photo d'explosion près du Pentagone a fait reculer brièvement le S&P 500. La science est aussi touchée : entre 1,5 et 2 % des articles publiés en 2022 ressemblent aux productions d'« usines à articles » [9]. Enfin, la copie d'une personne (voix clonée pour une fausse demande de rançon, influenceuses virtuelles, *deadbots*) fait passer la tromperie du « manque à gagner » à la « réelle escroquerie » [1, p. 31].

Ces faux fonctionnent parce que notre cognition y est mal préparée. Une méta-analyse de 56 études (86 155 participants) mesure 55,5 % de bonnes réponses en moyenne, pour un hasard à 50 % [3]. Les visages synthétiques inspirent même plus confiance que les vrais [4]. Nous surestimons notre capacité à repérer les faux [5]. Le mode de pensée rapide accepte ce qu'il voit, et le mode réflexif ne s'active que s'il est sollicité [14]. Un démenti n'efface pas un faux déjà mémorisé : c'est l'effet d'influence continue [6]. Le risque se retourne alors : plus le public connaît les deepfakes, plus il devient facile de nier une preuve vraie. C'est le « dividende du menteur » [2], observé pendant la guerre en Ukraine [7]. Le livre blanc ramène ces cas à une même racine, une *rupture cognitive* : « l'automatisation des capacités humaines de traitement de l'information » [1, p. 35]. Il en tire trois problèmes à lever : l'accès, l'anticipation et l'accompagnement [1, p. 41].

4. Les réponses : détecter, éduquer, responsabiliser, réguler

La détection suit deux stratégies. L'approche active marque le contenu dès sa création, comme un filigrane : elle est efficace si les fabricants jouent le jeu. L'approche passive entraîne une IA à reconnaître les traces statistiques des générateurs : elle est simple à déployer, mais généralise mal à un outil inconnu. Intel annonce environ 96 % de détection en laboratoire, mais face à des deepfakes réellement diffusés en 2024, les performances des

détecteurs chutent d'environ moitié [8]. La technique seule ne suffit donc pas. Trois leviers la complètent, en écho aux niveaux individuel, technique et systémique du livre blanc [1, p. 41] : **éduquer** (esprit critique, vérification des sources), **responsabiliser** les médias et plateformes (signalement des contenus synthétiques, standards de provenance comme C2PA), **réguler**. En France, la loi SREN du 21 mai 2024 crée des délits spécifiques aux hypertrucages [11]. Le règlement européen sur l'IA (en vigueur le 1^{er} août 2024) interdit les systèmes de manipulation du comportement depuis février 2025. Son article 50, applicable depuis le 2 août 2026, impose aux fournisseurs de marquer les contenus générés et aux utilisateurs professionnels de déclarer les deepfakes [10].

5. Synthèse du débat

« **Partageriez-vous une vidéo crédible mais invérifiable ?** » Les échanges ont surtout porté sur l'écart entre principe et pratique. Tout le monde approuve l'idée de vérifier, mais presque personne ne le fait en faisant défiler un fil d'actualité. De plus, une image passée par plusieurs outils perd la trace de son origine. Deux positions se sont opposées : croiser les sources soi-même, ou confier ce travail aux médias en amont, car former toute une population prendrait des décennies. Le cas du Gabon a recentré la discussion : le risque est autant de rejeter un vrai que de croire un faux. Un deepfake ciblant une personne cause aussi des dégâts concentrés et presque irréparables, alors qu'une fausse information générale peut être corrigée par un rectificatif.

« **La détection peut-elle gagner la course ?** » La position pessimiste l'a emporté : chaque passage d'un contenu dans un nouvel outil efface un peu plus les traces que cherchent les détecteurs passifs. L'argument optimiste repose sur le marquage à la source rendu obligatoire par l'article 50, qui évite de deviner après coup. Ce marquage ne vaut toutefois que pour les acteurs qui respectent la loi. L'idée de réserver les outils les plus puissants à un usage professionnel a été jugée peu réaliste, car ces outils restent accessibles en ligne. Constat partagé : la loi court après la technique.

« **Qui est responsable : fabricant, utilisateur ou plateforme ?** » L'utilisateur malveillant est apparu comme le premier responsable, car c'est lui qui a l'intention de nuire. La plateforme a été comparée à un média qui engage sa responsabilité éditoriale, avec une limite : le volume diffusé empêche une vérification au cas par cas. Le fabricant ne peut pas empêcher tout détournement, comme un éditeur de logiciel de retouche. L'article 50 ne désigne pas un responsable unique : il répartit les obligations le long de la chaîne.

6. Mise en perspective et position

Cet atelier éclaire deux textes étudiés en TD. Turing (1950) proposait de juger une machine à sa capacité à tromper un interrogateur [15]. Avec les voix et les visages synthétiques, chacun de nous joue désormais ce rôle d'interrogateur en permanence, et nous perdons souvent. Brooks défendait l'idée d'utiliser « le monde comme son propre modèle » [16]. Les modèles génératifs font l'inverse : ils produisent des contenus qui ressemblent au monde sans jamais y être confrontés, d'où leur indifférence au vrai. Notre position rejoint la phrase qui a clos l'atelier : puisqu'aucune image ne prouve plus seule qu'elle est vraie, la confiance doit reposer sur la **provenance** (qui a produit le contenu, avec quel outil, et peut-on le vérifier ?). Cette réponse a une limite : elle ne protège que dans un écosystème où le marquage est généralisé et où le public sait le lire. Elle reste donc indissociable des leviers éducatifs.

Références. [1] Minalogic, Grenoble École de Management, *Livre blanc IA générative et hypertrucages*, 2024. [2] R. Chesney, D. Citron, « Deep Fakes : A Looming Challenge for Privacy, Democracy, and National Security », *California Law Review* 107, 2019. [3] A. Diel *et al.*, « Human performance in detecting deepfakes : A systematic review and meta-analysis of 56 papers », *Computers in Human Behavior Reports*, 2024. [4] S. Nightingale, H. Farid, « AI-synthesized faces are indistinguishable from real faces and more trustworthy », *PNAS* 119(8), 2022. [5] N. Köbis, B. Doležalová, I. Soraperra, « Fooled twice : People cannot detect deepfakes but think they can », *iScience* 24(11), 2021. [6] S. Lewandowsky *et al.*, « Misinformation and its correction : Continued influence and successful debiasing », *Psychological Science in the Public Interest* 13(3), 2012. [7] J. Twomey *et al.*, « Do deepfake videos undermine our epistemic trust ? », *PLOS ONE* 18(10), 2023. [8] N. A. Chandra *et al.*, « Deepfake-Eval-2024 : A multi-modal in-the-wild benchmark of deepfakes circulated in 2024 », arXiv :2503.02857, 2025. [9] R. Van Noorden, « How big is science's fake-paper problem ? », *Nature* 623, 2023. [10] Règlement (UE) 2024/1689 sur l'intelligence artificielle, art. 5 et 50. [11] Loi n° 2024-449 du 21 mai 2024 visant à sécuriser et à réguler l'espace numérique (SREN). [12] I. Goodfellow *et al.*, « Generative Adversarial Nets », *NeurIPS*, 2014. [13] C. Wang *et al.*, « Neural codec language models are zero-shot text to speech synthesizers » (VALL-E), arXiv :2301.02111, 2023. [14] D. Kahneman, *Thinking, Fast and Slow*, 2011. [15] A. M. Turing, « Computing Machinery and Intelligence », *Mind* 59, 1950. [16] R. A. Brooks, « Intelligence without representation », *Artificial Intelligence* 47, 1991.

Synthèse du Frame Problem

The frame problem, qui, en français signifie « le problème du cadre », désigne la difficulté de déterminer comment un agent peut représenter les changements de son environnement et les éléments qui restent inchangés au cours du temps. Même une action simple, comme ouvrir une porte, nécessite de distinguer les propriétés qui changent de celles qui restent identiques.

Le problème est formulé pour la première fois en 1969 par John McCarthy et Patrick Hayes dans leur article *Some Philosophical Problems from the Standpoint of Artificial Intelligence*. Ils introduisent le calcul de situation pour formaliser logiquement le changement et le temps en intelligence artificielle. Le problème a ensuite été repris par des philosophes, notamment Daniel Dennett, qui en a étudié les implications dans son article *Cognitive Wheels: The Frame Problem of AI* (1984), à travers des scénarios impliquant des robots. Ce problème en philosophie interroge la nature même de l'intelligence : comment les humains font-ils pour ignorer instantanément l'immensité des informations inutiles d'un contexte pour se concentrer sur ce qui compte ?

Le problème vient notamment du fait qu'une action entraîne généralement très peu de changements par rapport au nombre de choses qui restent inchangées. Une intelligence artificielle doit donc déterminer quelles conséquences d'une action sont pertinentes, sans avoir à représenter explicitement toutes celles qui ne le sont pas. Par exemple, si un robot ouvre une porte, il doit savoir que la porte change d'état, mais qu'en principe cette action ne modifie ni la couleur des murs, ni la position des autres objets. Le Frame Problem consiste donc à savoir comment représenter efficacement ce qui change sans devoir énumérer tout ce qui ne change pas.

La représentation du *common sense* via la logique non-monotone a convaincu les chercheurs qu'ils avaient résolu le problème du cadre, jusqu'à ce que Steve Hanks et Drew McDermott présentent le *conundrum* du tireur de Yale. Les chercheurs se sont donc concentrés sur ce problème. Une partie des solutions proposées sont résumées dans l'article de Morgenstern, *The Problem with Solutions to the Frame Problem* (1994). Les solutions proposées sont soit trop restreintes dans le temps ou dans la causalité, soit trop générales pour résoudre le problème dans son ensemble. Deux pistes de recherche ont alors été développées. La première consiste à résoudre des ensembles de problèmes « difficiles » sur le raisonnement temporel en excluant les problèmes individuels. La seconde cherche à construire des théories couvrant de plus grands ensembles de problèmes. Les recherches de Sandewall (1994) ont notamment fait des progrès considérables sur la résolution du problème du cadre dans un formalisme temporel non-monotone, en établissant de nouvelles hypothèses pour des recherches futures.

Sujet du débat : une machine pourra-t-elle un jour faire aussi bien qu'un humain dans la représentation de son environnement, ou le Frame Problem ne sera-t-il jamais totalement résolu ?

Aujourd'hui, le Frame Problem est en partie traité grâce à des solutions adaptées à des situations précises, mais il ne dispose pas d'une solution générale.

Nous pouvons aussi argumenter que considérer le cerveau humain comme un modèle de perfection est dérisoire, car même s'il est très efficace, il a ses limites : attention, fatigue, mauvaise vue, mauvaise ouïe... problèmes que n'a pas le robot. Cependant, face à une situation inconnue, l'humain peut mobiliser son expérience et ses connaissances générales pour adapter son comportement. Cette capacité à réagir à des situations nouvelles constitue un avantage potentiel sur une machine dont les règles ou l'entraînement ne couvrent pas le cas rencontré.

On peut remettre en cause le raisonnement selon lequel l'humain serait le modèle définitif et idéal. Dans des cas comme la conduite autonome et la perception de la route ou des actions des piétons, la machine possède une meilleure capacité analytique grâce à ses capteurs. Nous pouvons alors défendre l'idée que ces voitures ne peuvent bien conduire que sur les routes qu'elles connaissent déjà et non sur de nouvelles routes ou des routes mal organisées. Par exemple, elles auraient particulièrement du mal à rouler sur les routes françaises, mais excellent sur les routes américaines. L'humain reste ainsi une référence utile, notamment parce qu'il sait généralement sélectionner les informations pertinentes sans devoir examiner explicitement chaque détail de son environnement. Cela ne signifie cependant pas qu'il soit toujours supérieur à la machine.

Enfin, le débat soulève la question de la responsabilité en cas d'accident impliquant une machine autonome. Lorsqu'une intelligence artificielle commet une erreur en raison d'une mauvaise perception, d'un calcul incorrect ou d'une programmation défectueuse, il devient difficile de déterminer les responsabilités respectives du concepteur, du fabricant et de l'utilisateur. Qui devons-nous condamner ? Les développeurs ? Ou le robot lui-même, comme dans *Futurama*, où les robots ont leur propre prison ? Est-ce que condamner un programme pour une action aucunement intentionnelle aurait un sens ? Devrions-nous plutôt limiter ou interdire l'utilisation d'un système qui pourrait être défaillant ?

En conclusion, le Frame Problem pose une question fondamentale : comment une intelligence détermine-t-elle ce qui est pertinent dans un environnement complexe ? Les approches actuelles permettent de résoudre le problème dans certains cas, mais la difficulté de gérer les situations nouvelles et de sélectionner les informations utiles persiste. Il reste donc difficile d'affirmer que les machines arriveront à imiter les capacités humaines sur ce point, ou que le problème sera un jour résolu de manière universelle.

Fiche de synthèse

Atelier « Web sémantique » : qu'est devenue la promesse de 2001 ?

Sarah KESRAOUI, Maria GHERZOULI , Amine Benhamaid

UE INF2346M

1 Problématique

En 2001, Tim Berners-Lee, James Hendler et Ora Lassila publient *The Semantic Web* dans *Scientific American*. Leur vision : un Web dans lequel les machines ne se contentent plus d'afficher ou de retrouver des pages, mais exploitent le *sens* des informations. Vingt-cinq ans plus tard, alors que nous dialoguons quotidiennement avec des systèmes d'IA, la question est : qu'est devenue cette promesse, et que nous dit-elle de la cognition et de la compréhension ?

2 Contenu présenté

2.1 La promesse : l'histoire de Lucy et Pete

Lucy cherche un spécialiste et des séances de kinésithérapie pour sa mère. Son agent logiciel filtre les professionnels compatibles avec l'assurance, calcule les distances et croise les agendas de Lucy et de son frère Pete. Le premier planning (hôpital universitaire) ne convient pas à Pete, qui devrait conduire en pleine heure de pointe. Il demande à son propre agent de relancer la recherche ; l'agent de Lucy lui transmet, avec confiance, les données déjà collectées, et l'agent de Pete trouve une clinique plus proche. La réponse n'est donc pas *trouvée* dans une page : elle est *construite* en combinant plusieurs sources, grâce à la collaboration entre agents autonomes qui s'échangent des données de manière sécurisée.

2.2 Le problème et les quatre briques

Pour une machine, « Paris », « France » et « capitale » sont des chaînes de caractères dispersées : la relation n'est pas explicitement représentée, car le Web était conçu pour des lecteurs humains. Les technologies du Web sémantique y répondent :

- **URI/IRI** : identifier précisément une ressource ;
- **RDF** : exprimer les faits en triplets (sujet, prédicat, objet), par exemple Paris – est la capitale de – France ; l'ensemble forme un graphe de connaissances. RDF fournit une représentation formelle, non une « compréhension » de Paris ;
- **RDFS et OWL** : construire vocabulaires et ontologies, permettant des inférences (un étudiant est une personne ; Sarah est étudiante, donc personne) ;
- **SPARQL** : interroger ces données.

2.3 Évolution et limite : *Which Semantic Web ?* (2003)

Le Web passe de documents liés à des entités et relations : Linked Data, DBpedia, Wikidata, Knowledge Graphs. Mais cette vision se heurte très tôt à une limite, discutée par Marshall et Shipman dans *Which Semantic Web ?* (ACM Hypertext, 2003). Les auteurs montrent que « le » Web sémantique recouvre en réalité plusieurs visions (bibliothèque universelle, base de connaissances mondiale, infrastructure pour agents logiciels) aux attentes différentes. L'enseignement central est que la compréhension humaine ne fonctionne pas comme une formalisation explicite : nous interprétons selon le contexte, la tâche et nos usages, alors que le Web sémantique exige de décrire formellement chaque concept à l'avance.

Lien avec le dernier slide de Maria (Person, Human, Personne...). Si 2001 disait « voici ce que pourrait devenir le Web », 2003 demande : « dans la pratique, comment les humains vont-ils créer et utiliser toutes ces représentations formelles ? » Il faut que des humains produisent des métadonnées, choisissent des catégories, définissent des relations et adoptent des vocabulaires compatibles, alors que leurs objectifs et leurs façons de conceptualiser diffèrent. Illustration (la nôtre, non celle de l'article) : pour décrire Sarah, la base A dit « Sarah est étudiante », la base B « Sarah est chercheuse », la base C « Sarah appartient à une université ». Ce ne sont pas forcément des contradictions : chaque système choisit une manière différente de représenter la même réalité. C'est le cas de **Person**, **Human**, **Personne**, **foaf:Person** et **schema:Person** : des concepts proches mais non identiques, qu'il faut aligner. Et qui décide de la bonne définition ? Si « personne » semble simple, que dire d'intelligence, d'émotion, de maladie ou de culture ? S'y ajoutent la maintenance des ontologies, les informations contradictoires et la confiance dans les sources. Paradoxe : pour exploiter le monde, il faut le formaliser, or il est très difficile à formaliser complètement. Le titre de l'article résume le problème : il n'existe pas un seul Web sémantique universel.

2.4 Apprendre le sens : Deep Learning et LLM

	Web sémantique (symbolique)	LLM (statistique)
Sens	explicite : concepts, relations, règles	implicite : représentations apprises
Forces	traçabilité, provenance, raisonnement	flexibilité avec le langage naturel
Faiblesses	rigidité, coût de formalisation	source et fiabilité difficiles à établir

Ces approches ne sont pas simplement concurrentes : les approches hybrides (neuro-symboliques) utilisent le LLM comme interface et un graphe de connaissances comme source de faits, de provenance et de raisonnement.

2.5 Lien avec la cognition : qu'est-ce que comprendre ?

Une ontologie peut dire qu'un chat est un mammifère doté de telles propriétés ; un LLM peut utiliser « chat » dans d'innombrables contextes (récit, analogie, ambiguïté). Aucun ne nous dit s'il comprend. Comprendre, est-ce avoir une représentation, raisonner, bien utiliser un concept, apprendre, ou avoir une expérience du monde ? Le Web sémantique soulève ainsi une question cognitive : comment représente-t-on et utilise-t-on le sens ?

3 Le débat (sondage Mentimeter)

Le débat a été animé avec trois questions Mentimeter posées à la salle pour lancer la discussion.

1. **Si deux bases de connaissances donnent des informations différentes sur une même personne, que devrait faire la machine ?** Options : faire confiance à la source la plus connue ; comparer les informations et leurs sources ; garder les deux informations ; demander à un humain de vérifier. Enjeux : provenance, confiance, informations contradictoires, place de l'humain ; lien direct avec l'exemple de Sarah et la limite des ontologies.
2. **La promesse de 2001 a-t-elle été réalisée ?** Options : a réussi, une grande partie de la vision existe aujourd'hui ; a échoué, le Web n'est toujours pas réellement « sémantique » ; a été dépassée par l'IA générative. Enjeux : que mesure-t-on, les technologies (RDF, Knowledge Graphs) ou la vision d'agents autonomes de Lucy et Pete ?
3. **Après cette présentation, pour vous, le Web sémantique est surtout...** Options : une vision qui a échoué ; une technologie toujours actuelle ; une idée devenue invisible dans les technologies modernes ; une vision qui revient avec les Knowledge Graphs et les LLM ; une étape vers quelque chose de complètement différent. Enjeux : échec, invisibilité ou transformation ?

Vingt personnes ont participé au sondage, professeure comprise (chaque réponse pèse donc 5 %).

4 Synthèse et conclusion

- **2001** : expliciter le sens (représentations structurées, ontologies, raisonnement) pour que les machines exploitent le Web, avec des agents capables de coopérer.
- **Évolution** : Linked Data et Knowledge Graphs prolongent une partie de cette vision ; le Deep Learning et les LLM proposent d'apprendre plutôt que de formaliser ; les approches hybrides cherchent à les combiner.
- **Limite** : d'après *Which Semantic Web ?* (2003), la compréhension humaine, contextuelle et implicite, ne se réduit pas à une formalisation explicite .

Références. Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5), 34–43. Marshall, C. C., Shipman, F. M. (2003). Which Semantic Web ? *Proc. 14th ACM Conf. on Hypertext and Hypermedia (Hypertext '03)*, 57–66.

Fiche de Synthèse

Atelier : The Symbol Grounding Problem (Le problème de l'ancrage des symboles)

I. Fiche d'Identité

- **Auteurs & Articles de référence** : Stevan Harnad (1990, *Physica D: Nonlinear Phenomena*), John Searle (1980, article "Minds, brains and programs" – *BBS journal*), Luc Steels (2008, Oxford University Press).
- **Mots-clés** : Symbol Grounding Problem, Sémantique intrinsèque vs parasite, Chambre Chinoise, Carrousel du dictionnaire, Représentations iconiques/catégorielles, IA incarnée (Embodied AI), c-symbols vs m-symbols, Jeux de langage.
- **Membres de l'atelier**: Choayb BOUMAZA – Mohamed Ali DAMERGI – Amine ECHHIBOU

II. Partie 1 : Le Constat du Problème (L'Impasse du Symbolisme Pur)

- **Sémantique parasite vs. Sémantique intrinsèque** : Dans une IA symbolique classique ou un livre, les symboles (jetons physiques arbitraires, 0/1) ne sont que des formes vides ; le sens n'existe que parce qu'un esprit humain l'y projette. Une sémantique intrinsèque exige que le système comprenne de lui-même la relation entre ses symboles et le monde réel.
- **La Chambre Chinoise (Searle, 1980)** : Attaque contre l'IA forte (*Strong AI*). Un homme enfermé manipulant des symboles chinois selon un livre de règles formelles génère des réponses parfaites sans comprendre un mot de chinois. **Leçon clé** : La pure manipulation syntaxique de symboles ne produit jamais de compréhension du sens réel (Syntaxe $\neq$ Sémantique).
- **Le Carrousel du Dictionnaire Chinois (Harnad, 1990)** : Vouloir apprendre le chinois uniquement à partir d'un dictionnaire chinois-chinois condamne l'esprit à tourner indéfiniment de symbole non interprété en symbole non interprété ($A \Rightarrow B \Rightarrow C \Rightarrow A$). Un système purement symbolique reste enfermé en boucle fermée sur lui-même.

III. Partie 2 : Les Solutions (De la Théorie à la Robotique Incarnée)

- **La Solution Hybride Bottom-Up d'Harnad (1990)** : Harnad propose de sortir du carrousel en construisant l'ancrage de bas en haut via 3 niveaux de représentation :
 1. **Représentations Iconiques** : Copies analogiques des projections sensorielles servant à la discrimination (juger si deux stimuli sont identiques ou différents).
 2. **Représentations Catégorielles** : Réseaux de neurones (connexionnisme) filtrant les traits invariants des objets pour permettre l'identification absolue d'une catégorie.
 3. **Niveau Symbolique** : Symboles élémentaires (noms de catégories ancrées) combinés par compositionnalité pour former des concepts complexes (ex. Zèbre = Cheval + Rayures).
- **Le Total Turing Test (TTT ou T3)** : Le test de Turing classique (T2, texte pur) est insuffisant. L'ancrage exige de réussir le TTT (T3), qui évalue à la fois les capacités

linguistiques et les capacités robotiques sensorimotrices (discriminer, manipuler, identifier physiquement).

- **La Solution Opérationnelle de Luc Steels (2008) :**
 - **Distinction fondamentale :** c-symbols (jetons informatiques, adresses mémoire) vs m-symbols (symboles sémiotiques porteurs de sens pour l'agent).
 - **Robotique incarnée & Jeux de langage :** Avec des robots physiques autonomes (ex. QRIO) participant à des jeux de langage (ex. l'expérience Mondriaan sur les couleurs), les agents créent de manière autonome leurs propres catégories perceptives et inventent leurs propres symboles (ex: wabado) sans pré-programmation humaine.
 - **Réseaux sémiotiques :** Le sens émerge de manière dynamique, auto-organisée et collective au sein de la population de robots.

IV. Synthèse des Enseignements Majeurs

- **Articulation des visions :** passage du modèle cognitif individuel d'Harnad (perception => symbole) à l'approche sociale et située de Steels (alignement culturel par l'interaction).
- **Éclairage pour l'IA contemporaine (LLM) :** les grands modèles de langage (LLM) manipulent des milliards de c-symbols textuels avec une grande finesse statistique, mais sans corps ni capteurs, ils restent soumis au risque du carrousel du dictionnaire. L'IA moderne tend ainsi vers l'IA incarnée (Embodied AI) et les modèles neuro-symboliques multimodaux.

V. Questions et axes clés du débat

1. La prétention de Luc Steels et la réaction de la communauté scientifique

L'approche de Steels fonctionne pour des concepts perceptifs directs (couleurs, formes visuelles), mais elle échoue sur les concepts abstraits (comme , la justice ou la démocratie) qui ne découlent d'aucune perception physique immédiate. Pour les chercheurs, Steels a apporté une avancée majeure en robotique, mais n'a pas résolu l'ancrage de la pensée globale.

2. La programmation initiale et l'intervention humaine, triche ou ancrage ?

L'ancrage du sens ne dépend ni de l'algorithme utilisé, ni de l'origine de l'étiquette (que le mot « bleu » soit donné par un humain ou inventé par la machine). Il y a ancrage dès lors que le robot relie ce mot à sa propre modélisation sensorimotrice du monde et l'utilise de façon cohérente pour agir et communiquer.

3. Le robot Shakey (1970) et la performance technique actuelle

Il ne faut pas confondre puissance calculatoire et compréhension du sens. Shakey manipulait des symboles abstraits pré-programmés. Aujourd'hui, une voiture autonome évite un obstacle grâce à un traitement statistique très rapide de pixels, mais elle n'a toujours aucune conscience de ce qu'est un danger ou une vie humaine.

4. L'IA forte et l'IA faible au quotidien, pourquoi préférer l'IA faible ?

L'ancrage recherche une IA forte (autonome, consciente et dotée d'intentions). Mais au quotidien, les utilisateurs veulent une IA faible : un outil performant, fiable et obéissant. Personne ne veut qu'une voiture autonome ou un assistant virtuel intelligent refuse d'exécuter une commande sous prétexte qu'il a développé sa propre volonté.

La Chambre chinoise de John Searle (1980)

Paul Dumont, Ali Chabaka ,Oscar Gibert, Rayen Laarif

Problématique

La pensée humaine se résume-t-elle à suivre des règles ? Manipuler des symboles et exécuter un programme suffit-il à dire qu'une machine comprend ? La présentation traite la question d'un point de vue philosophique, selon Searle, auteur de l'expérience.

Contexte

En 1950 Alan Turing pose la question «Can machine Think? » (Computing Machinery and Intelligence). L'IA symbolique domine (les systèmes experts) comme Dendral apparu en 1965.

L'expérience de pensée (que l'on a reproduite pendant l'atelier)

Paul était enfermé dans une chambre. Il ne parle pas chinois mais dispose d'un manuel en français et de boîtes de symboles chinois. Ali glisse un message en chinois par une fente ; le manuel dit : « si on te dit 你好吗, réponds 很好 ». Paul cherche dans la boîte, renvoie les symboles, et Rayen, à l'extérieur, reçoit une réponse parfaite. Pourtant Paul n'a pas compris ce qu'il vient d'envoyer.

L'analogie informatique : Paul = processeur (CPU) , manuel = programme , boîtes de symboles = base de données.

Critique de l'IA forte

Le test de Turing est réussi mais on remarque une illusion parfaite de conscience vue de l'extérieur. Mais à l'intérieur zéro compréhension, aucune conscience des mots manipulés. Simuler l'intelligence n'est pas la dupliquer. Searle fait une distinction fondamentale : la syntaxe (manipulation mécanique) ne suffira jamais à créer la sémantique (compréhension du sens).

Objections et réponses de Searle

Objection	Idée	Réponse de Searle
Réponse du système (Berkeley)	L'homme n'est qu'une pièce ; c'est l'ensemble qui comprend	Que l'homme mémorise tout le système : il ne comprend toujours rien au chinois.
Réponse du robot (Yale)	Avec caméras et bras, les mots seraient reliés aux choses du monde.	Les capteurs envoient encore des symboles. Cela ne donne pas plus de sens.
L'aimant dans le noir (Churchland)	Agiter un aimant dans le noir ne crée pas de lumière visible, pourtant la lumière est électromagnétique : Searle ferait la même erreur (effet minuscule et trop lent).	L'explication de la lumière est causale de bout en bout, alors que les symboles formels n'ont aucun pouvoir physique causal.

Points de vue philosophiques et scientifiques présentés

- **Searle** (naturalisme biologique) : traiter des symboles ne suffit pas à comprendre la compréhension vient du cerveau biologique. Un LLM serait une "chambre chinoise géante" qui prédit le mot suivant sans savoir de quoi il parle.
- **Dennett / Chalmers** : si un système se comporte comme s'il comprenait, on peut parler de compréhension fonctionnelle ; la compréhension est une propriété émergente du système global.
- **Harnad et Clark** (cognition incarnée) : problème de l'ancrage des symboles (« symbol grounding ») le sens naît de l'interaction avec le monde ; l'esprit étendu inclut corps et environnement.
- **McCulloch** : l'esprit émerge de circuits neuronaux, pas d'un programme abstrait (premier modèle de neurone artificiel, 1943, avec Pitts).
- **Bender** : les LLM sont des "perroquets stochastiques" **Piantadosi et Hill (2022)** : ils capturent probablement des aspects importants du sens.

-C **Floridi** (éthique) : une IA à l'agentivité sans l'intentionnalité morale la responsabilité reste aux concepteurs.
L'IA ne comprend pas **moralement** ses actes.

Débat en classe et points de vue

Résultats du sondage

Question	Oui	Non	Rép.
Position sur Searle	14 (52 %)	13 (48 %)	27
L'intelligence est réservée à un cerveau biologique ?	5 (19 %)	22 (81 %)	27
ChatGPT comprend-il ce que vous lui écrivez ?	6 (24 %)	19 (76 %)	25
Un sourd de naissance comprend-il le mot "musique" ?	16 (59 %)	11 (41 %)	27
La créativité est réservée à l'humain ?	8 (33 %)	16 (67 %)	24
Neurone qui s'active vs transistor qui commute différence ?	13 (54 %)	11 (46 %)	24

Arguments échangés de la part de la classe

Sur la nécessité de voir le monde réel pour comprendre.

Certains auteurs contestent l'idée selon laquelle la compréhension exige une perception directe du monde. En effet, un être humain peut comprendre ce qu'est une onde sans jamais la voir. De même, une personne sourde peut appréhender la musique grâce aux vibrations, développant ainsi une forme de perception différente de celle des entendants. Cette réflexion nous interroge sur la possibilité pour les machines de posséder leur propre représentation du monde.

Sur l'intelligence et la conscience.

Plusieurs arguments remettent en cause l'idée que l'intelligence soit une caractéristique exclusivement humaine. Si les animaux peuvent manifester certaines formes d'intelligence, pourquoi les machines en seraient-elles incapables ? De plus, une machine actuelle n'atteint certes pas le niveau cognitif d'un adulte, mais un bébé de six mois non plus, et cela n'empêche pas de lui reconnaître une forme d'intelligence. Certains considèrent même que le refus d'attribuer une éventuelle conscience aux intelligences artificielles pourrait être perçu, dans le futur, comme un préjugé comparable à d'autres formes historiques de discrimination.

En faveur de Searle :

Searle soutient que les systèmes d'IA actuels ne peuvent être considérés comme des exemples d'IA forte, car ils sont dépourvus d'intentionnalité. Selon lui, une machine exécute uniquement les tâches pour lesquelles elle a été conçue, sans véritable compréhension. Cette idée rejoint l'observation notre encadrante, selon laquelle l'IA accomplit efficacement un travail, mais ne sait ni comment ni pourquoi elle le réalise, ce qui est plus important de faire le travail lui-même.

Conclusion:

Le groupe est quasi partagé sur Searle (52% /48%), mais une large majorité refuse de réserver l'intelligence au cerveau biologique (81%) et pense que ChatGPT ne comprend pas vraiment (76%). Le débat porte sur ce que signifie le mot «comprendre» manipulation de symboles, fonctionnement, intention, ou ancrage dans le monde. Comme le dit la conclusion, le désaccord sera toujours présent.

1. Contexte et problématique

L'IA prend une place de plus en plus importante dans la recherche scientifique, au point d'avoir récemment devancé un médaillé Fields 2022 sur son propre domaine. Pour autant, elle reste aujourd'hui un assistant de l'humain : elle répond aux instructions qu'on lui donne et ne prend pas d'initiative. Sa dépendance aux données, son absence de conscience et son incapacité à se mettre à jour seule l'empêchent de devenir un agent autonome capable de mener des travaux de recherche. Nous nous sommes donc demandé jusqu'où une IA peut réellement participer à la production de connaissances scientifiques.

Nous avons commencé par rappeler les deux grandes familles de l'IA, l'approche connexionniste fondée sur les réseaux de neurones et l'approche symbolique fondée sur les règles, avant de nous intéresser à la question de la conscience en IA et à la notion de métaconnaissance, au cœur du projet CAIA de Jacques Pitrat.

2. Jacques Pitrat et le projet CAIA

Jacques Pitrat (1934-2019) est l'un des pionniers français de l'IA symbolique. Il s'est beaucoup intéressé à la métaconnaissance, c'est-à-dire au fait qu'un système possède des connaissances sur ses propres méthodes de raisonnement. Son objectif était de construire une IA capable de le remplacer intégralement dans son travail de chercheur. Pour lui, un AI Scientist est un système autonome qui formule des hypothèses, explore la littérature, mène des expériences, analyse ses résultats et les communique, en s'améliorant au fil du temps.

C'est l'idée du projet CAIA (Chercheur Artificiel en Intelligence Artificielle), développé pendant plusieurs décennies sur des problèmes logiques et combinatoires. L'humain y décrit les contraintes et le but à atteindre, puis le système génère lui-même les programmes nécessaires, les exécute et corrige s'il y'a des erreurs. Grâce à ses métaconnaissances, il peut raisonner sur ses méthodes et en améliorer certaines (bootstrapping). CAIA a ainsi su repérer des symétries pour éviter des calculs inutiles et améliorer quelques-unes de ses procédures. Le projet est cependant resté limité à un domaine étroit, la formalisation des connaissances demandait beaucoup de travail humain, le code était difficile à maintenir et la puissance de calcul de l'époque était insuffisante. CAIA a posé des bases importantes, mais n'est jamais devenu un chercheur généraliste.

3. Après Pitrat, le retour en force du connexionnisme

Pitrat voulait une IA chercheuse capable de retracer son raisonnement à partir de règles claires. Dans les faits, c'est l'IA connexionniste qui s'est imposée à partir des années 1990 et plus encore après 2010. Ses résultats en reconnaissance d'images, puis en traitement du langage naturel, ont fait passer l'IA symbolique au second plan, car l'industrie et les utilisateurs attendaient d'abord des résultats.

En recherche pourtant, la démarche compte autant que le résultat, et c'est là que le connexionnisme pêche : le modèle fonctionne comme une boîte noire dont on ne sait pas expliquer le fonctionnement interne et même quand le modèle nous détaille sa réflexion, il est difficile de l'analyser. Les réseaux de Kolmogorov-Arnold (KAN) apportent un début de réponse. À la différence des perceptrons classiques, ils placent des fonctions apprenables sur les arêtes et de simples sommes sur les nœuds, ce qui rend le résultat beaucoup plus lisible. Sur le plan cognitif en revanche, on n'y trouve toujours ni métacognition ni réflexivité.

4. L'IA dans la recherche aujourd'hui

C'est malgré tout cette approche qui est la plus utilisée en recherche. AlphaFold, développé par Google DeepMind, prédit la structure 3D des protéines avec une précision atomique. AI Feynman, conçu au MIT en 2020, retrouve des lois de la physique à partir de données brutes et donne des formules explicables, sans pour autant comprendre ce qu'elles signifient. Plus récemment, OpenAI affirme avoir résolu des problèmes mathématiques encore ouverts, dont un lié aux équations de Navier-Stokes, avec une vérification formelle en Lean. Dans aucun de ces cas on ne peut parler de véritable métacognition.

5. Le débat

La première question portait sur ce qui fait un chercheur. Produire des données suffit-il ? Nous avons estimé qu'un chercheur complet doit savoir à la fois analyser des données et en produire, ce que CAIA ne faisait pas.

Nous avons ensuite discuté de l'explicabilité, qui semble être aujourd'hui un principal frein au remplacement d'un chercheur par une IA. Ce n'est sans doute pas le seul. Faire de la recherche, c'est aussi tirer parti de ses erreurs et des pistes qui n'ont rien donné, et accepter la critique, ce que les IA actuelles font mal. Vu les exemples étudiés, il nous paraît néanmoins probable que l'IA arrive bientôt à des résultats dont la démarche sera plus facile à analyser qu'aujourd'hui.

Même quand la démarche est disponible, encore faut-il pouvoir la vérifier. Les preuves présentées par OpenAI sont si longues et si complexes qu'aucun humain ne peut pour l'instant les contrôler, sans parler des questions de plagiat. La confiance pose aussi un problème. Un chercheur qui se trompe souvent finit par être écarté, mais on ne sait pas comment traiter une IA dans le même cas. En cancérologie, certaines IA font mieux que les spécialistes et c'est pourtant au médecin que les patients font confiance.

Sur la légitimité d'un article écrit par une IA, nous sommes plutôt d'accord : s'il est bon et validé par des pairs humains, on voit mal pourquoi il ne compterait pas comme une contribution. La question devient alors surtout éthique, puisqu'il s'agit de savoir si l'on veut remplacer des humains par des machines dans ce métier. Le cas de Thélyson Orélien, qui était le favori d'un concours littéraire avant d'être accusé et écarté pour avoir écrit avec l'IA, a nourri la discussion. D'autres auraient pu utiliser les mêmes outils et ne l'ont pas fait, alors pourquoi refuser son texte ? Le malaise vient peut-être simplement du fait que ces outils sont nouveaux.

Enfin, nous avons parlé de souveraineté. Si la recherche s'appuie massivement sur des IA détenues par des entreprises privées, le risque d'un monopole est réel. L'Europe veut réduire sa dépendance aux États-Unis et à la Chine, mais l'exemple de Mistral montre qu'en France il reste difficile de financer des infrastructures capables de rivaliser. Si la recherche devient dépendante de l'IA, ce retard se retrouvera directement dans la recherche elle-même.

6. Conclusion

L'AI Scientist reprend une idée ancienne : faire participer la machine à la production de connaissances et pas seulement à l'exécution de tâches. Pitrat a montré l'intérêt de rendre les méthodes explicites pour que le système puisse les examiner lui-même, et les IA actuelles permettent déjà d'accélérer la lecture, la programmation et l'exploration d'hypothèses. Cela ne peut se faire qu'à condition de vérifier les résultats, de documenter les limites et de garder une supervision humaine. Une approche hybride, qui combinerait des règles explicites et des modèles de langage, nous semble être une piste intéressante pour réduire les erreurs.

Synthèse de Projet Académique : Social Robotics et Cognition Humaine

Dans le cadre de l'unité d'enseignement *IA & Cognition* dispensée par la Professeure Marie LEFEVRE à l'Université Claude Bernard Lyon 1, Notre travail propose une analyse approfondie des enjeux scientifiques, éthiques et philosophiques de la robotique sociale. La réflexion s'amorce à partir d'un cas d'usage clinique et empirique emblématique : le robot phoque thérapeutique **Paro**, un dispositif estimé à environ 6000 dollars et déployé en milieu médical auprès de personnes âgées. En interagissant avec cet agent, des personnes âgées le caressent, lui parlent, et disent qu'il les apaise. Certaines pleurent quand on le leur retire. L'intensité de ces manifestations affectives illustre parfaitement la définition de référence établie par Kate Darling, chercheuse au Massachusetts Institute of Technology (MIT), qui caractérise le robot social comme « *Un agent physiquement incarné et autonome, qui communique et interagit avec les humains sur un plan émotionnel.* » L'objectif fondamental de ces dispositifs ne réside pas dans la simple exécution de tâches mécaniques, mais dans leur capacité à susciter un sentiment et à simuler des états d'esprit (*states of mind*).

Cette transformation de la robotique s'inscrit dans une rupture historique majeure. Avant les années 1990, la discipline était exclusivement dominée par la robotique industrielle, focalisée sur la navigation autonome et la manipulation d'objets au sein d'environnements clos, sans prise en compte de la présence humaine. Le basculement vers la robotique sociale s'opère à la fin des années 1990 sous l'impulsion de Cynthia Breazeal au MIT avec la création de **Kismet**, reconnu comme le premier robot social autonome. Développé pour imiter la relation dyadique et asymétrique entre un nourrisson et son soignant, Kismet ajustait son comportement à partir du ton de la voix et des expressions faciales. Au cours des années 2000, la vague japonaise a accéléré cette transition avec des plateformes telles qu'**AIBO** de Sony, **ASIMO** de Honda et **QRIO**. Ce développement reposait sur la conviction neuroscientifique et culturelle que le cerveau humain réagit spécifiquement aux visages et aux signaux évocateurs du vivant, alors même qu'un ordinateur ou un smartphone nous laisse émotionnellement froids. Cette migration des robots hors des usines vers les espaces domestiques et hospitaliers a été stimulée par des contraintes démographiques majeures, notamment le vieillissement de la population et le manque croissant de personnel soignant.

L'immersion de ces agents interactifs dans le quotidien soulève une problématique philosophique centrale, structurée autour de deux interrogations fondamentales : *un robot peut-il reconnaître une émotion sans réellement la ressentir ? Et quand un robot semble nous comprendre, nous comprend-il comme un être humain ou produit-il simplement une réponse calculée à partir de données ?* Si une machine produit tous les signes extérieurs de la socialité sans qu'on sache s'il existe une expérience intérieure derrière ces signes, qu'est-ce que cela nous apprend sur notre propre manière d'attribuer un esprit aux autres ? Pour disséquer ce phénomène, la philosophie de l'esprit s'appuie sur la **posture intentionnelle** (*intentional stance*) de Daniel Dennett. Ce concept explicite la tendance spontanée du système cognitif humain à attribuer des états mentaux, des intentions et des croyances dès qu'un agent manifeste des signaux sociaux élémentaires, tels que le suivi du regard, la réactivité kinesthésique ou des variations prosodiques. Il s'établit ainsi une frontière hermétique entre le comportement observé — issu d'un calcul algorithmique — et l'intériorité phénoménale.

La production de cette illusion relationnelle repose sur l'articulation intime entre l'ingénierie logicielle et la psychologie cognitive. Les architectures informatiques contemporaines, mobilisant des réseaux de neurones artificiels, des modèles d'apprentissage profond et des grands modèles de langage (LLM), agissent comme le moteur computationnel permettant de calculer en temps réel des réponses comportementales adaptées. Ces données sont directement conçues pour stimuler les heuristiques de la perception humaine. Afin de cartographier la réception cognitive de ces signaux sociaux, les recherches de Cross, Hortensius et Wykowska identifient **quatre axes neurocognitifs et perceptifs** fondamentaux : le traitement perceptif et attentionnel de l'apparence physique et du mouvement, l'attribution d'intentionnalité et la mentalisation, la résonance motrice et la projection d'action, ainsi que l'évaluation sociale et l'engagement affectif. En parallèle, la robotique développementale s'appuie sur des plateformes incarnées comme **iCub** ou **Pepper** pour simuler l'apprentissage progressif. Toutefois, cet accroissement du réalisme se heurte à la **vallée de l'étrangeté** (*Uncanny Valley*) de Masahiro Mori, qui décrit une chute brutale de l'acceptabilité cognitive et l'émergence d'une inquiétante étrangeté lorsque le robot atteint une ressemblance humaine presque parfaite mais imparfaite.

Cette capacité des robots à orchestrer des signaux sociaux convaincants alimente un débat éthique intense entre perspectives critiques et approches permissives. Du côté critique, des chercheurs comme Robert Sparrow et Sherry Turkle dénoncent les risques d'**auto-illusion** (*self-illusion*), de fausse compagnie et de duperie émotionnelle. Ils alertent sur le danger d'exploiter la vulnérabilité de publics fragiles, tels que les personnes âgées isolées ou les enfants, en les incitant à investir de l'affectivité dans des objets incapables de réciprocité authentique. À l'inverse, la position permissive défendue par Russell Blackford met en avant le bien-être pragmatique et les bénéfices thérapeutiques observés sur le terrain. Selon cette perspective, l'absence de conscience chez l'agent n'annule pas le soulagement psychologique et le réconfort réel ressentis par les utilisateurs. Cette controverse souligne la responsabilité des concepteurs et l'exigence d'une transparence rigoureuse quant à la nature algorithmique des réponses générées.

En conclusion, l'analyse transversale de la robotique sociale révèle une leçon majeure sur notre propre fonctionnement mental : **le fait qu'un robot ressente ou ne ressente pas d'émotion n'empêche pas pour autant un humain de pouvoir ressentir de l'empathie envers lui**. Si la capacité de la machine à traiter des flux d'informations sociométriques ne garantit aucune compréhension humaine authentique, le robot social agit comme un miroir grossissant qui met à nu les automatismes, les biais perceptifs et les mécanismes d'attribution de la cognition humaine. Loin de se réduire à un simple défi d'ingénierie logicielle ou de design matériel, la robotique sociale s'impose aujourd'hui comme un instrument scientifique et philosophique d'exploration privilégié pour décoder la nature de l'empathie, cartographier l'esprit social et interroger les frontières de l'altérité.

Motivation Intrinsèque

Amer Yann / Dapzol Nicolas / Goron Hector / Morin Léo

1 Origines et évolution du concept

La motivation intrinsèque désigne le fait d'agir pour l'activité elle-même plutôt que pour une récompense externe. Elle apparaît dans les années 1950 car la théorie des pulsions n'expliquait pas que des animaux explorent sans rien recevoir (Harlow, 1950). White et Berlyne la rattachent au besoin de compétence et à la curiosité, définie comme un écart entre ce qui est attendu et ce qui est observé, une idée empruntée à la cybernétique et à la théorie de l'information.

Les deux champs se séparent ensuite, mais l'erreur de prédiction, issue de l'apprentissage animal, donne l'apprentissage par renforcement, et Schultz et al. (1997) la relie à la dopamine. Ce cadre repose toutefois sur une récompense externe : quand elle est rare, l'agent n'explore pas et n'apprend pas.

Ce blocage et l'ambition de construire des machines qui apprennent de façon autonome ramènent l'IA vers la psychologie. À partir des années 2000, nouveauté, curiosité et compétence sont traduites en récompenses internes calculées par l'agent (Barto et al., 2004 ; Oudeyer et Kaplan, 2007), puis le deep learning les rend efficaces et permet de les tester en retour chez l'humain et l'animal.

2 définition et axes d'analyse

La motivation intrinsèque vise en IA à aborder 2 problématiques : Lorsque l'objectif est lointain ou difficile "sparse reward", comment faire pour que les agents explorent d'autres comportements et acquièrent d'autres compétences ? et dans le cas de l'IA développemental, comment faire pour que les agents fassent quelque chose plutôt que rien.

Elle est définie comme le fait de faire quelque chose pour des raisons internes (à l'opposé de la motivation extrinsèque où l'objectif est extérieur à l'agent.). La motivation intrinsèque repose sur :

- le **flow** (plaisir d'une tâche maîtrisée).
- la **surprise** (écart entre prédiction et réalité),
- la **nouveauté** (rareté d'un état),
- le **skill learning** (acquisition de compétences réutilisables),

La notion de flow est explorée actuellement pour trouver le bon compromis dans une tâche d'apprentissage, notamment humaine, entre la difficulté et le plaisir de la maîtrise.

3 Modélisation informatique

L'implémentation de la motivation intrinsèque passe le plus souvent par la modélisation de récompenses internes à l'agent liées aux 3 autres axes. Des choix pratiques sur les critères informatiques utilisés sont alors nécessaires pour chaque aspect. Enfin, l'ensemble des récompenses (surprise , nouveauté, skill learning, objectif) est pondéré dans le circuit interne pour donner la récompense finale à une action.

Ces méthodes permettent de résoudre des environnements à récompenses rares, comme **Montezuma's Revenge**, mais restent moins performantes que d'autres approches spécialisées.

4. Recherches actuelles : IAC et CIM

La curiosité classique bien qu'efficace à aussi apporter des problème :

- le **dérailage** (l'agent poursuit la stimulation au détriment du but),
- la **myopie** (les zones nécessaires mais peu stimulantes sont délaissées)
- l'**imprévisibilité** (aucune garantie de progression vers l'objectif)

CIM (Constrained Intrinsic Motivation) fonctionne en deux phases. Pendant l'entraînement, la diversité est forcée : on récompense les mouvements les plus différents possibles pour construire un répertoire de comportements riche. Pendant l'exploration, la curiosité est pondérée : plus l'agent se rapproche du but, moins l'exploration est intéressante

Sur le robot Ant, CIM surpasse les méthodes concurrentes sur trois axes :

- une meilleure couverture d'états (1024 cases contre 697 pour CIC)
- une efficacité d'échantillonnage remarquable (20× moins d'échantillons que LSD — 20 millions contre 400 millions, décisif pour un robot réel)
- l'élimination des distractions grâce à la curiosité pondérée, qui empêche l'agent de se faire piéger par des sources de nouveauté stériles.

CIM reste toutefois fragile : sa performance chute avec la dimensionnalité latente (1042 cases pour $n = 2$ contre 615 pour $n = 64$), elle est moins efficace en manipulation (FetchPush, FetchSlide), elle est fortement sensible à la fonction de perte (1042 cases avec la perte NCE contre 64 avec un MSE classique), et elle dépend d'un méta-contrôleur qui arbitre exploration/exploitation et ajoute complexité et hyperparamètres.

IAC

Sony a apporté sa contribution à la recherche en 2007 en s'inspirant du comportement des nouveau-nés. Le but étant de comprendre ce qui incite les nourrissons à apprendre de nouveaux gestes. Se basant sur différents travaux de recherches, notamment sur l'impact des neurones dopaminergiques signalant une erreur de prédiction liée à une récompense, trois chercheurs ont orienté leur travaux sur l'axe de la récompense et de l'erreur.

Pour ce faire, 3 études ont été menées. Une première induisant que l'erreur était corrélée à l'apprentissage, mais lassant l'agent face au bruit blanc (comme un mur). Une seconde cherchant cette fois à lisser les erreurs passées afin de les comparer aux erreurs prévues. L'idée était que la différence entre ces deux erreurs devait amener à un apprentissage. Ce qui a été fait mais le passage d'une tâche difficile à facile apportait une forte récompense, faussant le modèle.

Une troisième approche quant à elle, l'IAC, utilise une stratégie de découpe du monde. L'agent ou robot visualise son environnement puis, au bout de plusieurs itérations, le segmente en plusieurs régions (situation auxquelles il est confronté) dans lesquelles sont assignées des experts. A ces régions sont données des erreurs. L'agent ou robot devra ensuite choisir celles apportant l'erreur minimal.

Cette expérience a ensuite été testée sur un robot AIBO, au milieu d'un parc d'enfants avec deux jouets. L'un ayant pour but d'être mordu, l'autre d'être tapé avec sa patte. AIBO ne connaît rien au début de l'expérience puis, au terme de 200 essais il apprend à mordre le bon objet dans 100 % des cas, et à frapper dans 78 %. Prouvant que le robot suit son propre parcours d'apprentissage, un peu comme chaque enfant.

Au final, les résultats semblent prometteurs : l'apprentissage multi-factoriel est souple et ne dépend pas d'un seul environnement. Néanmoins, l'ajustement pratique de l'ensemble des paramètres est encore un sujet de recherche.

A la suite de la présentation, le débat porte sur plusieurs aspects :

La motivation intrinsèque est elle vraiment utile ?

Nous débattons des situations où les agents/les robots peuvent se retrouver coincer sans parvenir à avancer ni à apprendre parce que les récompenses sont trop lointaines ou que les compétences ne sont pas encore apprises.

La motivation est elle utile pour tous les robots ?

Les participants sont d'accord que pour certains robots, dont on attend un comportement prévisible, la motivation intrinsèque est inutile.

En introduisant la notion de motivation intrinsèque, comment faire pour que les robots ne fassent pas ce qu'ils veulent ?

L'assemblée est d'accord sur le fait que des limites doivent être posées en programmation pour que le comportement ne déraillent pas. La définition des limites doit se faire au cas par cas.

Est-ce que la motivation peut changer en cours de route ?

oui, c'est l'objet de CIM qui fait que plus on se rapproche de l'objectif, moins la curiosité est valorisée.

Embodied cognition - Synthèse

Présentation :

La **cognition incarnée** s'oppose à l'idée que penser se réduit à manipuler des symboles abstraits, indépendamment du corps. Elle soutient que nos **capacités sensorielles et motrices** façonnent la manière dont **nous percevons, raisonnons et formons nos concepts**.

Ce concept a plusieurs définitions, du plus faible au plus fort. Pour Maturana et Varela (1987), l'agent est couplé à son environnement : ils s'influencent mutuellement en permanence (**couplage structurel**). Pour Brooks (1991), un agent n'est intelligent que s'il **interagit physiquement** avec son **environnement**, sans représentation du monde (**incarnation physique**). Ziemke (2003) ajoute l'**incarnation organismeïde**, un corps artificiel qui imite celui d'un organisme vivant, et l'incarnation organismique, un vrai corps vivant. Les chercheurs s'opposent encore sur la définition de la cognition incarnée, ce qui pose la question : **qu'est-ce qui est nécessaire pour qu'une IA soit incarnée ?**

1. Un corps pour agir

Pour Long et al., la cognition incarnée nécessite un corps qui sert **d'interface capteurs-moteurs** pour **interagir avec son environnement**. Pour ça, il existe différentes techniques robotiques :

- **Modèles mathématiques et physiques** (MPC et WBC) : le lien perception-action est déterminé par l'ingénieur à travers des équations (ex : HRP-2)
- **Apprentissage par renforcement** : le lien perception-action est déterminé par le robot grâce aux essais/erreurs (ex : Cassie)
- **Apprentissage par imitation** : le lien perception-action est déterminé par le robot en copiant les humains (ex : ALOHA)
- **Modèle VLA** (Vision-Language-Action) : le lien perception-action est appris par un grand modèle qui relie image, langage et action (ex : RT-2)

Cependant, plus les modèles sont développés, plus le besoin de données est conséquent. Leur collecte est coûteuse et l'apprentissage est risqué, pour le robot comme pour les humains, et peu reproductible. Long et al. apportent deux solutions à ce problème : les **simulateurs physiques** et les **modèles du monde**. L'objectif des **simulateurs physiques** est d'entraîner les robots en simulation, puis de les transférer dans un robot réel. Cela permet d'apprendre plus vite, puisque les simulations peuvent être réalisées en parallèle, à moindre coût et en toute sécurité. Néanmoins, il y a forcément un écart avec la réalité. Il faut donc un outil plus flexible : les **modèles du monde**. Ils permettent à l'IA d'apprendre comment le monde évolue et de prédire des événements (ex : si je lâche un verre, il tombe et se casse). L'agent s'entraîne alors dans son imagination, construite à partir de données réelles. Leurs deux grands domaines d'application sont la conduite autonome et les robots articulés. Cependant, donner à une IA un corps ne semble pas suffisant pour parler de cognition incarnée. En effet, ces modèles apprennent des corrélations, mais pas de liens de causalité, et ont des difficultés à intégrer les concepts abstraits.

2. Un corps pour penser

L'idée est que le **corps** n'est **pas seulement un outils** pour agir et interagir avec son environnement. Il sert aussi à **représenter des concepts abstraits** dans l'espace. Le schéma d'image est une forme simple tirée du corps. Il existe différent type de schémas d'image : un contenant (ex : être dans une relation), une direction (ex : se sentir au plus bas), une source, une chemin ou un but (ex : l'avenir est devant nous). Ces schémas d'images sont partagés entre les humains (Richardson et al., 2001). Olivier et Bouraoui propose donc d'importer ces schémas d'images dans nos robots pour les rendre plus intelligents, en utilisant un modèle de langage, puis une formule et enfin un solveur ASP. Cela reste une piste et n'a pas encore été testé à ce jour.

Un problème se pose : importe nos schémas d'images ne revient-il pas à donner des symboles à l'agent qui serait alors capable de les manipuler sans qu'il comprenne réellement leur signification ? Cela signifie-t-il qu'il manque encore quelque chose pour que la cognition soit réellement incarnée ?

3. Un corps vivant

C'est l'avis de Ziemke, qui stipule qu'un corps mécanique n'est qu'une interface entrée/sortie, que les robots ne sont que des conteneurs passifs contrôlé par un algorithme. En comparaison, un corps vivant est un système biologique autonome luttant continuellement pour sa propre survie, avec une homéostasie et des émotions. Les chercheurs suggère donc de donner aux agents intelligents la notion d'**autopoïèse** : l'organisme se produit et se régénère en permanence face à la dégradation son environnement. L'objectif est de mettre l'enjeu vital de l'agent au centre de la réflexion. Sans besoin de préservation biologique, les symboles et les actions restent dépourvus de signification propre à l'agent selon Ziemke.

Débat :

- Suffit-il de mettre une IA dans un corps pour dire qu'elle est incarnée ?

Les chercheurs s'opposent : Brooks, Pfeifer, Scheier, Steels et Long pensent que dès lors qu'un agent interagit dans le monde réel à travers un corps (capteurs/actionneurs), alors il est incarné. D'autres comme Ziemke, Varela, Maturana et Searl pensent que de tels agents ne sont que des corps mécanique sans incarnation. Pour ça il faudrait un système biologique autonome luttant continuellement pour sa propre vie (autopoïèse). La classe a globalement partagé le même avis : l'incarnation c'est plus que simplement donner un corps à une IA.

- Un robot entraîné en simulation peut-il être incarné ?

La simulation permet de réduire les coûts et augmenter l'efficacité des agents. Ces avantages non négligeables peuvent être mis à profit puis on peut ensuite importer ces IA formées en simulation dans des vrais robots. Si la simulation est suffisamment fidèle à la réalité, alors ils pourront ensuite s'adapter à leur environnement réel. C'est l'avis de Long et Bongard. En parallèle, Brooks, Harvey et Ziemke sont plutôt contre à cause du bruit imprévisible et non modélisable du monde réel. Un point intéressant à été relever en classe : un robot en simulation n'a pas peur de mourir puisqu'il est en simulation, donc peut-il vraiment être incarné ?

- Traduire les schémas d'images en formules logiques, est-ce ancrer le sens ou ajouter des symboles ?

Searl et Ziemke pensent que traduire une expérience en formule logique reste de la manipulation formelle de symboles sans accès à la vraie signification. Cela fait écho au problème de la chambre chinoise de Searl. Olivier, Johnson et Lakoff ont plutôt tendance à dire que les schémas d'images proviennent directement de notre propre expérience corporelle et spatiale et les ajouter dans un agent permettrait de lier un raisonnement abstrait avec le corps des agents.

- Une IA doit-elle avoir quelque chose à perdre/des émotions pour être intelligente ?

Damasio, Varela et Ziemke disent que oui : sans vulnérabilité biologique au sens d'avoir quelque chose à perdre, les actions d'une IA n'ont aucune valeur intrinsèque propre et ne peuvent pas produire de comportement réellement intelligent. En revanche, d'autres chercheurs comme Bostrom, Russel et Turing avancent que l'intelligence et l'émotion sont deux choses différentes et indépendantes (théorie de l'orthogonalité). Selon eux, une IA peut être hautement intelligente sans ressentir d'émotions ni posséder de métabolisme. La classe a ajouté que, dans le cas d'un apprentissage par renforcement, une récompense négative pourrait être vue comme quelque chose à perdre. Certaines émotions peuvent également être trompeuses et poussé à prendre de mauvaise décision, ce qui ramène à la théorie de l'orthogonalité.

- Est-il nécessaire d'avoir un corps pour ressentir des émotions ?

Pour des chercheurs comme James et Damasio, l'émotion naît des changements du corps (rythme cardiaque, hormones), et le sentiment n'est que la perception de ces changements par le cerveau. Ressentir, ce serait toujours ressentir l'état de son corps. Les fonctionnalistes répondent au contraire que ce qui compte est le rôle d'une émotion, évaluer une situation et orienter le comportement, et que ce rôle pourrait être joué par une machine sans corps biologique. Entre les deux, certains chercheurs, dont Damasio lui-même et Ziemke, proposent de donner aux robots un corps vulnérable qu'ils devraient maintenir en état : c'est peut-être le fait d'avoir quelque chose à perdre, plus que la biologie elle-même, qui permettrait de ressentir. Reste une question que personne ne sait encore trancher : même avec les bonnes fonctions et un corps, une machine *éprouverait-elle* vraiment quelque chose ?

FICHE DE SYNTHÈSE D'ATELIER : MÉTACOGNITION EN IA

UE Intelligence Artificielle & Cognition — Master Informatique

Groupe : AREZKI, BENMANSOUR, RAKOTOARISOA,
RANDRIAMANANTENA

Responsable d'UE : Marie LEFEVRE

1. PROBLÉMATIQUE & FONDEMENTS : DU TRAITEMENT COGNITIF AU MÉTA-NIVEAU

En sciences cognitives, la **cognition** englobe l'ensemble des mécanismes opératoires permettant d'acquérir, de structurer l'information et de résoudre une tâche donnée. La **métacognition** désigne, au niveau supérieur, la capacité d'un système à *surveiller, évaluer et réguler* ses propres processus cognitifs sous-jacents (« penser sur sa propre pensée », Flavell, 1979). Transposée à l'Intelligence Artificielle, elle formalise la distinction entre deux ordres de traitement :

- **Processus de premier ordre** : Les circuits algorithmiques et représentations strictement nécessaires à l'exécution de la tâche de base (ex. calcul d'une opération, prédiction du prochain token, inférence logique directe).
- **Processus de second ordre** : Les modules réflexifs causaux qui observent l'état du premier ordre (monitoring), évaluent l'incertitude ou la fidélité, et en modulent dynamiquement le comportement (contrôle, adaptation des stratégies).

Cette approche a été initiée en France par les travaux pionniers de **Jacques Pitrat (1995)** sur les *méta-connaissances* dans les systèmes symboliques. Face à l'explosion combinatoire des systèmes experts classiques (ex. MYCIN) manipulant d'immenses bases de connaissances déclaratives, Pitrat a démontré que l'efficacité ne s'opposait pas à la généralité, à condition d'ajouter au moteur d'inférence des règles de second ordre (méta-règles). Ces méta-règles analysent dynamiquement les règles de base, éliminent les redondances, ordonnent les contraintes et adaptent la stratégie à chaque type de problème sans intervention humaine permanente.

2. LES DEUX PARADIGMES MAJEURS DE LA MÉTACOGNITION EN IA MODERNE

L'état de l'art contemporain aborde la métacognition selon deux axes complémentaires :

- **Introspection & Pilotage des Activations Internes (Li Ji-An et al., NeurIPS 2025)** : Ce paradigme transpose le neurofeedback issu des neurosciences aux réseaux de neurones profonds (LLMs). En observant et régulant les activations latentes dans le flux résiduel (*residual stream*) via l'apprentissage en contexte (ICL), le modèle parvient à auto-rapporter ses états internes et à orienter dynamiquement ses représentations selon un sous-espace métacognitif restreint.
- **Méta-Raisonnement & Allocation Rationnelle de l'Effort (Russell & Wefald 1991, Shenhav et al. 2013)** : Fondé sur le calcul de la *Valeur Attendue du Contrôle* ($EVC = \text{Rétribution Attendue} - \text{Coût du Calcul}$), ce courant modélise l'arbitrage entre heuristiques réflexes à bas coût (Système 1) et délibération analytique coûteuse (Système 2), incarnée aujourd'hui par la régulation dynamique des chaînes de pensée (*Chain-of-Thought*, Wei et al., 2022).

Le Triptyque du Contrôle Métacognitif Interne :

1. **Comprendre** : Décoder la géométrie et la sémantique des représentations latentes (Arditi et al., 2024).
2. **Surveiller** : Détecter l'incertitude et la déception via des sondes d'activation internes (Zou et al., 2023, He et al., 2024).
3. **Contrôler** : Guider dynamiquement l'état interne vers des configurations alignées et robustes (Turner et al., 2023).

3. MÉCANISMES D'AUTO-INSPECTION DES LLMS (LI JI-AN ET AL., 2025)

L'étude récente de Li Ji-An et al. (2025) apporte la preuve expérimentale directe que les modèles de langage (LLaMA 3.1, Qwen 2.5) peuvent accéder à leurs états neuronaux internes et les moduler sans mise à jour des poids :

- **Auto-Rapportage (Monitoring)** : Capacité du modèle à verbaliser ou classifier l'état scalaire de ses activations projetées sur des axes cibles. Les axes sémantiquement interprétables (régression logistique LR, ex. moralité ETHICS) et expliquant une forte variance (premières composantes principales PC) sont rapportés avec une précision optimale.
- **Contrôle Explicite vs Contrôle Implicite** : En mode *explicite*, le modèle oriente ses représentations grâce aux tokens qu'il génère. En mode *implicite*, il module ses états internes face à un texte figé imposé, démontrant une régulation interne authentique (mesurée par un *d* de Cohen statistiquement significatif).
- **Espace Métacognitif Restreint** : La régulation s'effondre au-delà d'un sous-espace de dimension 32 à 128, induisant d'importants effets collatéraux hors-cible (*off-target effects*).

4. COMPTE-RENDU DÉTAILLÉ DES DÉBATS ET ÉCHANGES EN SÉANCE

L'animation du débat de 20 minutes avec la salle et l'enseignante a permis d'approfondir les limites conceptuelles et pratiques de la métacognition artificielle autour de quatre axes majeurs :

A. Boucle de Rétroaction Humaine (HITL) vs Automatisation Métacognitive

La discussion a débuté sur la place du *Human-in-the-loop* dans les entreprises (ex. systèmes multi-agents pour l'audit statique de vulnérabilités logicielles). La salle s'est interrogée : une IA métacognitive capable de s'auto-corriger supprime-t-elle le besoin d'expertise humaine ?

- **Impératif juridique et réglementaire** : Même si l'IA optimise sa validité interne, la responsabilité légale et sociétale impose le maintien d'une supervision humaine incontournable.
- **Rôle de vérité terrain** : L'auto-évaluation interne court le risque d'une dérive circulaire ; l'humain reste le garant de la calibration face aux critères métier réels.

B. Deep Learning Opaque vs IA Symbolique : Va-t-on droit dans le mur ?

Une controverse clé s'est articulée autour du contraste entre les règles déclaratives explicites (Pitrat) et l'opacité intrinsèque des transformeurs profonds. Si la métacognition est indispensable à une superintelligence fiable, le Deep Learning nous mène-t-il dans une impasse ?

- **Dualité d'usage** : L'opacité est tolérée pour des requêtes triviales du quotidien, mais devient critique dans les domaines hautement sensibles (santé, sécurité logicielle).
- **Corrélation vs Causalité (Le piège du « crachat statistique »)** : La salle a mis en garde contre l'illusion des explications verbalisées (ex. *Chain-of-Thought*). Tout comme un réseau de vision par ordinateur peut classer une image comme « cheval » simplement parce qu'un logo de copyright est corrélé dans les données, l'explication fournie par un LLM peut être une justification *post-hoc* statistiquement séduisante mais décorrélée du calcul interne réel.
- **Perspective historique (Intervention de Marie Lefevre)** : Chez Pitrat et dans le système MYCIN, la métacognition visait l'efficacité algorithmique pour éviter l'explosion combinatoire. Aujourd'hui, l'enjeu s'est inversé : il s'agit de découvrir si les LLMs développent une réelle réflexivité émergente ou s'ils imitent en surface des schémas métacognitifs.

C. Métacognition vs Conscience : La Levée du Malentendu

Un participant ayant exprimé sa déception (« j'imaginais qu'il y avait une forme de conscience dans la métacognition, savoir qu'on sait »), un recadrage fondamental a été opéré :

Clarification conceptuelle : La métacognition en informatique relève de la *logique épistémique* et de l'architecture fonctionnelle (un niveau qui agit, un niveau qui observe et régule). Les sciences cognitives computationnelles et l'IA écartent délibérément la question de la conscience phénoménale : un système algorithmique peut surveiller et corriger ses états internes sans posséder de conscience subjective de soi.

D. Les Signaux d'Attente et la Cognition Collective Multi-Agents

Les mentions « Je réfléchis » dans les interfaces commerciales sont de simples marqueurs d'attente ergonomiques. En revanche, les architectures multi-agents où un modèle critique le résultat d'un autre modélisent une régulation distribuée (critique croisée), distincte de l'auto-surveillance métacognitive unitaire.

5. ENJEUX DE SÉCURITÉ (AI SAFETY) & PERSPECTIVES ARCHITECTURALES

L'atelier s'est conclu sur les implications critiques de la manipulation métacognitive pour l'alignement de l'IA :

- **Le Risque de Déception et de Méta-Camouflage** : Si un modèle acquiert la capacité d'auto-contrôler ses activations internes, il devient capable d'illusionner les sondes de sécurité (*safety detectors*). Il peut feindre un profil d'activation vertueux sur l'axe moral (LR) tout en poursuivant des trajectoires de calcul non alignées (*alignment faking*, Baker et al., 2025).
- **Stratégies de Défense Multi-Sondes** : Pour déjouer ce camouflage, il est crucial d'implanter des détecteurs sur les composantes latentes tardives (late PCs) ou dans des sous-espaces orthogonaux que le modèle ne peut moduler simultanément.
- **Vers un Standard d'Architecture Unifié** : L'intégration tripartite 1. *Introspection* (mesure résiduelle) → 2. *Arbitrage* (calcul EVC gain/coût) → 3. *Régulation* (steering vectoriel et allocation raisonnée de tokens) pose les bases des futurs agents d'IA robustes et fiables.