Création de Systèmes d'IA Agentiques : Des Architectures Classiques aux Paradigmes Fondés sur les Grands Modèles de Langage

Sous-titre : Une Analyse des Fondements Théoriques, des Implémentations Pratiques et des Implications Socio-Techniques

Auteur : Dr. A. Chevalier

Université : Université Sorbonne AI

Département : Département d'Informatique et d'Intelligence Artificielle

Date : 30 juillet 2025

Résumé

Ce mémoire aborde la problématique de la création de systèmes d'intelligence artificielle (IA) agentiques, un domaine en pleine expansion qui marque une transition fondamentale des IA réactives vers des entités autonomes, proactives et capables d'interagir avec leur environnement pour atteindre des objectifs complexes. La problématique centrale de cette recherche est d'analyser comment la convergence des architectures d'agents symboliques classiques, notamment le modèle Croyance-Désir-Intention (BDI), et des capacités de raisonnement sub-symboliques des grands modèles de langage (LLM) redéfinit les principes de conception, les capacités fonctionnelles et les défis de gouvernance des systèmes d'IA autonomes.

La méthodologie adoptée est une analyse de synthèse qualitative et critique, s'appuyant sur une revue de littérature exhaustive des travaux fondateurs en systèmes multi-agents (SMA), des architectures cognitives, de l'apprentissage par renforcement multi-agent (MARL) et des publications les plus récentes sur les agents basés sur les LLM. Cette approche conceptuelle et comparative permet de jeter un

pont entre deux traditions de l'IA souvent perçues comme distinctes.

Les principaux résultats de cette étude démontrent que, loin d'être obsolètes, les concepts fondamentaux de l'agentivité (autonomie, réactivité, proactivité, sociabilité) et les cadres architecturaux comme le BDI connaissent une véritable "réincarnation" conceptuelle. Ils offrent des structures de contrôle et d'interprétabilité indispensables pour maîtriser le comportement des agents modernes. Nous établissons que la véritable autonomie des agents basés sur les LLM n'émerge pas du modèle de langage seul, mais d'une architecture hybride qui l'augmente avec des composants essentiels pour la planification, la mémoire à long terme (via des techniques comme le RAG) et l'interaction avec le monde réel (via l'utilisation d'outils). Cette architecture agentique est conceptualisée comme un "système d'exploitation" pour l'intelligence du LLM.

En conclusion, ce mémoire soutient que l'avenir de la création d'agents robustes et fiables réside dans une synthèse dialectique des approches symboliques et connexionnistes. Nous formulons des recommandations pour la recherche future, axées sur le développement de méthodes de vérification formelle pour ces systèmes hybrides et sur l'étude approfondie de l'alignement des valeurs dans des sociétés d'agents décentralisées. Pour les praticiens, nous soulignons l'importance d'intégrer les principes de conception des architectures classiques pour construire des agents basés sur les LLM plus prévisibles, contrôlables et dignes de confiance.

Table des Matières

(Générée automatiquement)

Liste des Tableaux, Figures et Acronymes

(Générée automatiquement)

Introduction Générale

Contexte et Justification du Sujet

Le domaine de l'intelligence artificielle (IA) connaît une transformation paradigmatique. Après des décennies de développement axées sur des systèmes spécialisés dans des tâches uniques, l'émergence de modèles capables d'une plus grande généralité a ouvert la voie à une nouvelle frontière : l'IA agentique. Cette évolution marque le passage de systèmes purement réactifs, qui se contentent de répondre à des commandes spécifiques, à des entités computationnelles proactives et autonomes, capables de percevoir leur environnement, de raisonner, de planifier et d'agir de manière indépendante pour atteindre des objectifs complexes.[1] L'IA agentique n'est plus un concept confiné aux laboratoires de recherche ou aux récits de science-fiction ; elle est devenue une réalité technologique tangible, dont l'impact se propage à travers tous les domaines de la société.[2]

Cette transition est catalysée par deux avancées majeures. La première est la maturité des recherches sur les systèmes multi-agents (SMA), qui ont fourni pendant des décennies les fondements théoriques de l'autonomie, de la coordination et de la communication décentralisée. La seconde, plus récente et spectaculaire, est l'avènement des grands modèles de langage (LLM). Ces modèles, dotés de capacités de compréhension et de génération de langage naturel sans précédent, agissent désormais comme de véritables "moteurs cognitifs", conférant aux agents des capacités de raisonnement et de planification d'une flexibilité jusqu'alors inégalée.

L'importance stratégique de ce domaine est attestée par les investissements massifs et les initiatives gouvernementales. En France, par exemple, la stratégie nationale pour l'IA, initiée en 2018 et renforcée par des investissements considérables, a positionné le pays comme un écosystème de premier plan en Europe, notamment dans le développement de modèles fondamentaux puissants.[2] Cet effort national et international reflète la conviction que l'IA agentique est un accélérateur d'innovation capable de transformer radicalement des secteurs clés. Dans la santé, les agents autonomes promettent d'améliorer la précision des diagnostics et de personnaliser les traitements.[3] Dans la finance, ils permettent une analyse des marchés en temps réel et une prise de décision automatisée.[4] La robotique, le transport autonome, l'éducation

personnalisée et la recherche scientifique elle-même sont également à l'aube de révolutions menées par ces nouvelles formes d'intelligence artificielle.[2]

Problématique

Malgré l'enthousiasme et les progrès rapides, le domaine de la création de systèmes d'IA agentiques est confronté à une fragmentation conceptuelle. D'une part, l'héritage de l'IA symbolique, incarné par des architectures délibératives robustes comme le modèle Croyance-Désir-Intention (BDI), offre des cadres formels pour un raisonnement explicite et un comportement orienté vers des buts. D'autre part, la vague de l'IA connexionniste, menée par les LLM, propose une approche sub-symbolique, flexible et puissante, mais souvent opaque et difficile à contrôler. La simple juxtaposition de ces deux paradigmes ne suffit pas à créer des agents fiables et dignes de confiance. Une véritable synthèse est nécessaire, mais ses principes restent à définir.

La problématique centrale de ce mémoire est donc formulée comme suit : **Comment la convergence des architectures d'agents symboliques classiques (telles que BDI) et des capacités de raisonnement sub-symboliques des grands modèles de langage (LLM) redéfinit-elle les principes de conception, les capacités et les défis de gouvernance des systèmes d'IA autonomes?**

Cette question fondamentale explore la tension et la synergie potentielles entre deux traditions historiques de l'intelligence artificielle. Elle cherche à déterminer si les modèles classiques sont rendus obsolètes par les LLM ou s'ils fournissent, au contraire, les échafaudages nécessaires pour structurer et maîtriser la puissance des nouvelles technologies.

Questions de Recherche et Hypothèses

Pour aborder cette problématique, nous avons formulé quatre questions de recherche principales :

1. **QR1 :** Quels sont les principes fondamentaux des architectures d'agents classiques qui restent pertinents pour la conception des agents modernes basés

sur les LLM?
2. **QR2 :** De quelle manière les LLM agissent-ils comme des "moteurs cognitifs" au sein des architectures agentiques, et quels sont les composants (mémoire, outils) indispensables pour traduire leur potentiel de raisonnement en action effective?
3. **QR3 :** Comment les défis de la coordination et de l'apprentissage dans les systèmes multi-agents (SMA) sont-ils exacerbés ou résolus par l'intégration des LLM?
4. **QR4 :** Quelles sont les implications de cette nouvelle vague d'autonomie agentique sur les plans éthique, économique et sociétal, notamment en matière de responsabilité, d'alignement des valeurs et de transformation du travail?

Ces questions de recherche sont guidées par deux hypothèses centrales :

- **H1 :** Loin d'être obsolètes, les concepts des architectures classiques comme BDI (Belief-Desire-Intention) offrent des cadres structurants essentiels pour maîtriser, interpréter et vérifier le comportement délibératif des agents basés sur les LLM, fournissant une sémantique de l'action qui fait défaut aux modèles de langage seuls.
- **H2 :** La véritable autonomie agentique ne découle pas du LLM seul, mais d'une architecture hybride qui intègre le raisonnement linguistique du modèle avec des modules externes spécialisés pour la perception, la mémoire à long terme et l'interaction effective avec l'environnement numérique ou physique (Tool Use).

Objectifs et Structure du Mémoire

L'objectif principal de ce mémoire est de fournir une analyse exhaustive et critique de la création de systèmes d'IA agentiques. Pour ce faire, nous visons à :

1. Cartographier l'évolution historique et conceptuelle du domaine de l'agentivité.
2. Synthétiser et comparer les paradigmes architecturaux, des modèles symboliques classiques aux approches modernes basées sur les LLM.
3. Évaluer les défis techniques fondamentaux liés à la conception d'agents uniques et de systèmes multi-agents, notamment en matière d'apprentissage, de coordination et de vérification.
4. Débattre des implications profondes de l'autonomie agentique sur la société, l'économie et l'éthique.

Pour atteindre ces objectifs, ce mémoire est structuré en trois parties distinctes.

La Partie I : Fondements Théoriques de l'Agentivité se consacre à l'établissement du cadre conceptuel. Le Chapitre 1 propose une revue de littérature approfondie sur la notion d'agent intelligent, en retraçant ses origines et en définissant ses propriétés fondamentales.

La Partie II : Architectures et Mécanismes de l'Intelligence Autonome constitue le cœur technique de notre analyse. Le Chapitre 2 examine en détail les architectures d'agents classiques, avec un focus particulier sur le modèle BDI. Le Chapitre 3 analyse la révolution apportée par les agents fondés sur les LLM, en décortiquant leurs composants architecturaux. Le Chapitre 4 étend l'analyse à la dynamique complexe des systèmes multi-agents. Enfin, le Chapitre 5 aborde les questions cruciales de la fiabilité, de la vérification et de l'évaluation de ces systèmes.

La Partie III : Implications et Perspectives élargit le champ de l'analyse aux applications et aux enjeux sociétaux. Le Chapitre 6 présente des études de cas concrètes dans divers secteurs. Le Chapitre 7 est dédié à une discussion critique des enjeux éthiques, économiques et sociétaux, et inclut une autocritique des approches actuelles.

Le mémoire se conclut par une discussion générale qui synthétise les apports de la recherche, suivie d'une conclusion qui récapitule les résultats et ouvre sur de nouvelles pistes de recherche.

Partie I : Fondements Théoriques de l'Agentivité

Chapitre 1 : Revue de la Littérature sur la Notion d'Agent Intelligent

Ce chapitre établit les fondations conceptuelles nécessaires à la compréhension des systèmes d'IA agentiques. Il retrace l'émergence historique de la notion d'agent, définit rigoureusement ses propriétés fondamentales, analyse les travaux pionniers qui ont structuré le domaine, et le distingue de paradigmes informatiques connexes.

L'objectif est de démontrer que les concepts classiques, loin d'être caducs, fournissent un cadre indispensable pour interpréter les avancées contemporaines.

1.1. Les Origines Conceptuelles : de l'IA Distribuée aux Systèmes Multi-Agents (SMA)

Le concept d'agent n'est pas né avec la récente vague des grands modèles de langage. Ses racines plongent dans le domaine de l'Intelligence Artificielle Distribuée (IAD), une branche de l'IA qui a émergé dans les années 1980.[6] L'IAD partait du postulat que l'intelligence n'est pas nécessairement une faculté centralisée et monolithique, mais peut émerger de l'interaction d'entités de calcul décentralisées. L'objectif était de résoudre des problèmes trop vastes ou trop complexes pour un seul processeur ou un seul système intelligent, en les décomposant en sous-problèmes pouvant être traités en parallèle par une collection d'entités.[7]

Au sein de l'IAD, deux courants principaux se sont développés. Le premier, la résolution de problèmes distribuée (*Distributed Problem Solving*), se concentrait sur la manière de décomposer et d'allouer des tâches à des modules coopératifs. Le second, qui nous intéresse plus particulièrement, a donné naissance aux Systèmes Multi-Agents (SMA). L'idée centrale des SMA est de modéliser des systèmes complexes comme des sociétés d'entités autonomes, appelées "agents", qui interagissent dans un environnement commun. Chaque agent est conçu avec des informations ou des capacités de résolution de problèmes limitées, lui conférant ainsi un point de vue partiel sur le système global.[7] Il n'existe pas de contrôle global centralisé ; le comportement du système dans son ensemble émerge des interactions locales entre les agents.[7]

Cette approche s'est avérée particulièrement puissante pour modéliser et simuler des systèmes complexes et dynamiques où les données et le contrôle sont naturellement décentralisés.[9] Des revues de littérature systématiques sur le domaine des SMA montrent une croissance exponentielle de la recherche avant 2009, suivie d'une spécialisation dans des domaines d'application spécifiques comme l'ingénierie des systèmes électriques ou la modélisation de réseaux complexes.[7] La recherche s'est également concentrée sur l'identification de "design patterns" pour les SMA, cherchant à capitaliser sur des solutions éprouvées pour des problèmes récurrents, bien que ce domaine souffre d'un manque de standardisation.[11] L'héritage de l'IAD et des SMA est donc fondamental : il a fourni le paradigme de la décentralisation, de

l'interaction et de l'émergence, qui constitue la toile de fond sur laquelle se déploient aujourd'hui les systèmes d'IA agentiques.

1.2. Définition et Propriétés Fondamentales de l'Agentivité

Sur la base de cet héritage, une définition consensuelle de l'agent intelligent a émergé dans la littérature. Un agent est un système informatique qui est **situé** dans un environnement (physique ou virtuel) et qui est capable d'**action autonome** dans cet environnement afin d'atteindre ses objectifs.[7] Cette définition, bien que simple, repose sur un ensemble de propriétés fondamentales qui distinguent un agent d'un simple programme. Les travaux fondateurs de Wooldridge et Jennings (1995) ont été particulièrement influents dans la formalisation de ces propriétés, qui restent la pierre angulaire de la définition de l'agentivité aujourd'hui.[13]

Ces quatre propriétés canoniques sont :

- **Autonomie :** C'est la caractéristique la plus essentielle. Un agent est autonome car il peut opérer sans l'intervention directe et constante d'un humain ou d'un autre agent. Il possède un contrôle sur ses propres actions et sur son état interne.[1] Cette capacité à fonctionner sans supervision humaine constante est ce qui différencie fondamentalement un agent d'un simple outil.[5]
- **Réactivité :** Un agent doit être capable de percevoir son environnement et de répondre de manière opportune (*in a timely fashion*) aux changements qui s'y produisent.[7] Cette perception peut se faire via des capteurs physiques dans un robot, ou via des API et des flux de données pour un agent logiciel. La réactivité implique que l'agent n'agit pas seulement selon des plans rigides, mais peut s'adapter aux dynamiques de son environnement.[1]
- **Proactivité :** L'agentivité ne se limite pas à la simple réaction. Les agents sont proactifs, ce qui signifie qu'ils ne se contentent pas de répondre à leur environnement, mais sont capables de prendre l'initiative pour atteindre leurs objectifs. Ils exhibent un comportement orienté vers des buts (*goal-directed behavior*).[1] Par exemple, un agent de voyage ne se contente pas de réserver un vol sur demande ; il peut anticiper des retards, proposer des itinéraires alternatifs et réserver un hôtel de manière proactive.[1]
- **Sociabilité :** Les agents opèrent rarement de manière isolée. Ils doivent être capables d'interagir avec d'autres entités, qu'il s'agisse d'autres agents logiciels ou d'utilisateurs humains. Cette interaction se fait généralement par le biais d'un

langage de communication d'agent (voir Chapitre 4), qui leur permet de coopérer, de se coordonner et de négocier.[7]

Ces quatre propriétés ne sont pas mutuellement exclusives ; un agent intelligent robuste doit les posséder toutes à des degrés divers. Elles forment un cadre conceptuel qui permet de passer de l'idée d'un programme qui exécute des instructions à celle d'une entité qui *agit* de manière intentionnelle dans un monde complexe.

1.3. Analyse Critique des Travaux Fondateurs

Le milieu des années 1990 a été une période charnière pour la formalisation du concept d'agent. L'article séminal de Wooldridge et Jennings, "Intelligent agents: theory and practice" (1995), a joué un rôle clé en synthétisant les diverses recherches de l'époque et en proposant un cadre unifié.[14] Leur travail a clairement articulé la distinction entre la théorie des agents (la formalisation mathématique de leurs propriétés), leurs architectures (les plans de conception logicielle) et les langages de programmation d'agents.[13] Cette distinction a structuré le champ de recherche pour les années à venir.

Parallèlement, d'autres chercheurs comme Nwana ont proposé des typologies pour classer la diversité croissante des systèmes d'agents. Nwana a identifié sept types d'agents : collaboratifs, d'interface, mobiles, d'information/Internet, réactifs, hybrides et intelligents (*smart agents*).[19] Cette taxonomie, bien que datée, met en lumière les différentes facettes de l'agentivité et l'idée que ces types peuvent être vus comme des dimensions dans un espace de conception, permettant la création de systèmes hétérogènes. Par exemple, un agent d'assistance personnel pourrait être à la fois un agent d'interface (interagissant avec l'utilisateur), un agent d'information (cherchant des données sur le web) et un agent collaboratif (négociant avec les agents d'autres utilisateurs pour planifier une réunion).[19]

Ces travaux fondateurs ont établi un programme de recherche ambitieux : construire des entités logicielles qui ne sont pas de simples automates, mais des systèmes capables de raisonnement pratique, de prise de décision autonome et d'interaction sociale complexe. Cependant, une critique que l'on peut formuler rétrospectivement est que les implémentations de l'époque, basées sur une IA symbolique et une planification logique rigide, peinaient à atteindre la flexibilité et la robustesse promises

par la théorie, en particulier dans des environnements ouverts et imprévisibles comme l'Internet. La vision était claire, mais les outils pour la réaliser pleinement manquaient encore.

C'est ici que se situe un point de bascule conceptuel. Les propriétés fondamentales de l'agentivité, définies dans les années 90, ne sont pas devenues obsolètes avec l'avènement des LLM. Au contraire, elles connaissent une sorte de "réincarnation". Les LLM fournissent un mécanisme sub-symbolique d'une puissance inédite pour *réaliser* ces propriétés à une échelle et avec une flexibilité auparavant inaccessibles. La proactivité, par exemple, était classiquement implémentée via des systèmes de planification symbolique, souvent fragiles et coûteux en calcul. Une IA générative standard comme ChatGPT est, par nature, purement réactive : elle répond à une instruction et s'arrête.[1] Cependant, une architecture agentique moderne qui place un LLM dans une boucle de perception-raisonnement-action transforme cette capacité de génération de langage en un comportement véritablement proactif et autonome.[1] En décomposant un objectif de haut niveau en sous-tâches, en utilisant des outils pour les exécuter et en s'auto-corrigeant en fonction des résultats, le système réalise la vision originale des pionniers du domaine, mais avec des moyens techniques radicalement différents. La véritable innovation n'est donc pas l'abandon des concepts classiques, mais leur accomplissement grâce à une nouvelle technologie. Le LLM n'est pas un agent en soi ; il est le composant qui permet de construire des agents qui satisfont enfin pleinement les définitions exigeantes de l'agentivité.

1.4. Distinction avec les Paradigmes Connexes

Pour bien cerner la spécificité de l'IA agentique, il est crucial de la distinguer de concepts avec lesquels elle est souvent confondue.

- **IA Agentique vs. IA Générative :** La distinction la plus importante à l'ère moderne est celle entre l'IA agentique et l'IA générative. Une IA générative, comme ChatGPT ou DALL-E, est un système **réactif**. Elle crée du contenu (texte, image, code) en réponse à une instruction humaine (*prompt*) spécifique. Son rôle se limite à la génération de cette sortie. Une IA agentique, en revanche, est **proactive**. Elle ne se contente pas de générer une réponse ; elle prend des initiatives, exécute une séquence d'actions dans son environnement et poursuit des objectifs de manière autonome sur la durée. L'IA générative est un composant, le "moteur cognitif", tandis que l'IA agentique est l'architecture

complète qui utilise ce moteur pour agir.[1]

- **Agent vs. Objet :** La distinction avec la Programmation Orientée Objet (POO) est plus ancienne mais tout aussi fondamentale. Les objets, au sens de la POO, sont des entités logicielles qui encapsulent un état et un comportement (via des méthodes). Cependant, ils sont fondamentalement **passifs** : leur comportement n'est activé que par la réception d'un appel de méthode externe. Les agents, à l'inverse, sont **autonomes** et possèdent leur propre fil d'exécution (*thread of control*). Ils peuvent décider de manière proactive quand et comment agir, sans attendre une sollicitation externe. La communication entre agents se fait par des messages asynchrones, ce qui renforce leur autonomie, alors que les appels de méthode en POO sont généralement synchrones.[7]

1.5. Identification des Lacunes et des Axes de Recherche

Cette revue de la littérature met en évidence une lacune majeure dans la recherche actuelle. Alors que les fondations théoriques de l'agentivité, issues de l'IA symbolique et des SMA, sont bien établies, et que les implémentations pratiques basées sur les LLM se multiplient à un rythme effréné, il manque un cadre conceptuel unifié qui relie ces deux mondes. La littérature contemporaine est souvent très axée sur l'ingénierie et les frameworks pratiques (par exemple, comment construire un agent avec LangChain), sans toujours reconnecter ces nouvelles architectures aux principes théoriques de la délibération, de l'intentionnalité ou de la coordination formelle qui ont été étudiés pendant des décennies.[1]

Par conséquent, un axe de recherche central, que ce mémoire se propose d'explorer, est la construction d'un pont conceptuel entre les architectures classiques (comme BDI) et les architectures modernes (basées sur les LLM). Comprendre comment les capacités de raisonnement flexible des LLM peuvent être intégrées dans les structures de raisonnement pratique éprouvées des agents symboliques est essentiel pour concevoir la prochaine génération de systèmes d'IA autonomes, qui devront être non seulement performants, mais aussi robustes, explicables et alignés avec des objectifs humains.

Partie II : Architectures et Mécanismes de l'Intelligence

Autonome

Chapitre 2 : Analyse des Architectures d'Agents Classiques

Avant l'avènement des grands modèles de langage, le champ de l'intelligence artificielle agentique s'est structuré autour de plusieurs paradigmes architecturaux distincts. Ces architectures, développées pour répondre à différents types de problèmes et d'environnements, représentent les fondations sur lesquelles reposent de nombreux concepts encore pertinents aujourd'hui. Ce chapitre analyse les trois principales familles d'architectures classiques : les agents réactifs, les agents délibératifs (avec un focus sur le modèle BDI), et les architectures hybrides qui tentent de combiner les avantages des deux premières.

2.1. Les Agents Réactifs et la Subsomption

Les architectures réactives représentent la forme la plus simple d'agentivité. Elles sont fondées sur un principe de couplage direct et rapide entre la perception et l'action, en évitant délibérément toute forme de raisonnement symbolique complexe ou de modélisation explicite du monde.[7] L'idée centrale est que l'intelligence peut émerger de l'interaction de comportements simples avec un environnement complexe, plutôt que d'une représentation interne détaillée de cet environnement.

L'exemple le plus emblématique de cette approche est l'**architecture de subsomption** (*subsumption architecture*), proposée par Rodney Brooks dans les années 1980 pour la robotique mobile. Cette architecture est organisée en une hiérarchie de couches de comportements. Chaque couche implémente une compétence spécifique, comme "éviter les obstacles" ou "se déplacer vers un point". Les couches supérieures peuvent subsumer (c'est-à-dire inhiber ou modifier) les sorties des couches inférieures. Par exemple, la couche "explorer" peut guider le robot, mais si la couche inférieure "éviter les obstacles" détecte un mur, elle prendra le contrôle pour empêcher une collision. Ce mécanisme permet une réponse rapide et

robuste aux événements imprévus de l'environnement.

La principale force des agents réactifs réside dans leur efficacité et leur vitesse de réaction dans des environnements dynamiques. Cependant, leur critique fondamentale est leur manque de vision à long terme. Comme ils ne possèdent pas de modèle explicite de leur environnement et ne planifient pas leurs actions, il leur est difficile de prendre des décisions qui nécessitent d'anticiper les conséquences futures ou de raisonner sur des objectifs complexes.[7] De plus, l'apprentissage à partir de l'expérience est un défi majeur pour les agents purement réactifs, car il n'y a pas de structure interne pour stocker et généraliser les connaissances acquises.[7]

2.2. Les Agents Délibératifs et Symboliques : le Modèle BDI

En opposition directe aux agents réactifs, les architectures délibératives postulent qu'un comportement intelligent nécessite une représentation symbolique explicite du monde et un processus de raisonnement pour décider de l'action à entreprendre. L'archétype de l'agent délibératif est celui basé sur le modèle **Croyance-Désir-Intention (BDI)**, qui est devenu un standard de facto dans la recherche sur les agents rationnels.[12]

2.2.1. Origines Philosophiques

Le modèle BDI n'est pas une simple construction informatique ; il est profondément ancré dans la philosophie de l'action, et plus particulièrement dans la théorie de la **raison pratique** (*practical reasoning*) du philosophe Michael Bratman.[25] Bratman cherchait à expliquer comment les humains prennent des décisions et agissent de manière rationnelle. Il a mis en évidence le rôle crucial des

intentions comme des états mentaux distincts des simples désirs. Un désir est un état du monde que l'on aimerait voir se réaliser, tandis qu'une intention est un désir auquel on s'est **engagé** (*committed*). Cet engagement a deux conséquences majeures : il nous pousse à élaborer des plans pour réaliser l'intention et il nous confère une certaine persistance dans nos actions, nous empêchant de reconsidérer nos objectifs à chaque instant.[27] C'est cette structure de raisonnement que l'architecture BDI

cherche à modéliser informatiquement.

2.2.2. Composants de l'Architecture BDI

L'architecture BDI est structurée autour de trois composants mentaux principaux, qui modélisent l'état interne de l'agent :

- **Croyances (Beliefs) :** Elles représentent l'état informationnel de l'agent, c'est-à-dire ce qu'il tient pour vrai concernant son environnement, lui-même et les autres agents. Il est important de noter qu'il s'agit de "croyances" et non de "connaissances", car elles peuvent être incomplètes, incertaines ou même fausses.[7] Dans la plupart des implémentations, les croyances sont stockées dans une base de faits, souvent sous forme de littéraux logiques.[26]
- **Désirs (Desires) :** Ils représentent l'état motivationnel de l'agent. Ce sont les objectifs ou les situations que l'agent aimerait accomplir ou voir se réaliser. Les désirs peuvent être multiples, potentiellement contradictoires, et pas nécessairement réalisables à un instant donné.[7] Dans de nombreuses implémentations, le terme "buts" (*goals*) est utilisé de manière interchangeable avec "désirs".
- **Intentions :** Elles représentent l'état délibératif de l'agent. Une intention est un sous-ensemble des désirs que l'agent a choisi de poursuivre activement. Cet acte de "choisir" implique un engagement. L'agent sélectionne un plan d'action pour réaliser son intention et s'y tient, à moins qu'une raison impérieuse ne l'oblige à reconsidérer cet engagement (par exemple, si l'intention est atteinte, si elle devient irréalisable, ou si un objectif plus prioritaire apparaît).[19]

2.2.3. Le Cycle de Raisonnement BDI

Le comportement d'un agent BDI n'est pas statique ; il est gouverné par un processus continu appelé le **cycle de raisonnement** ou l'**interpréteur BDI**. Ce cycle, qui s'exécute en boucle, réalise les deux étapes du raisonnement pratique identifiées par Bratman : la délibération et le raisonnement moyens-fins.[26] Bien que les détails varient d'une implémentation à l'autre, le cycle typique se déroule comme suit [26] :

1. **Mise à jour des Croyances :** L'agent observe son environnement (via ses

capteurs ou des messages reçus) et met à jour sa base de croyances pour refléter les changements perçus.
2. **Génération d'Options (Délibération) :** Sur la base de ses nouvelles croyances et de ses désirs, l'agent génère un ensemble d'options ou de désirs potentiellement réalisables.
3. **Filtrage (Sélection des Intentions) :** L'agent filtre ces options pour sélectionner un sous-ensemble cohérent et réalisable d'intentions auxquelles il va s'engager. Cette étape cruciale de délibération stratégique décide de "quoi faire".
4. **Planification (Raisonnement Moyens-Fins) :** Pour chaque intention nouvellement adoptée, l'agent sélectionne dans sa bibliothèque de plans une séquence d'actions appropriée pour l'atteindre. Cette étape de planification tactique décide de "comment le faire".
5. **Exécution :** L'agent exécute la prochaine action du plan associé à l'une de ses intentions actives. L'exécution de cette action modifie l'environnement, et le cycle recommence.

Le modèle BDI offre un cadre puissant pour construire des agents capables de raisonnement complexe et de comportement rationnel. Cependant, sa principale faiblesse est le risque d'être trop lent dans des environnements très dynamiques, car le processus de délibération peut être coûteux en temps de calcul.

2.3. Les Architectures Hybrides : Synthèse de la Réactivité et de la Délibération

Face aux limites respectives des approches purement réactives et purement délibératives, les chercheurs ont rapidement proposé des **architectures hybrides** pour combiner le meilleur des deux mondes.[12] L'idée la plus courante est une architecture en couches (

layered architecture).

Dans une telle architecture, plusieurs sous-systèmes de raisonnement coexistent et interagissent. Typiquement, on trouve :

- Une **couche réactive** à la base, responsable des réponses rapides et instinctives aux événements immédiats de l'environnement. Cette couche assure la survie et la réactivité de l'agent.
- Une **couche délibérative** au sommet, souvent basée sur le modèle BDI, qui s'occupe de la planification à long terme, de la prise de décision stratégique et de

la gestion des objectifs.
- Parfois, une **couche de coordination** intermédiaire qui gère les interactions entre la couche réactive et la couche délibérative, en décidant quelle couche doit avoir le contrôle à un instant donné.

Les architectures hybrides sont conceptuellement très séduisantes car elles miment la dualité du raisonnement humain (réflexe vs. réfléchi). Elles permettent à un agent d'éviter un obstacle de manière réactive tout en continuant à planifier son itinéraire pour atteindre une destination lointaine. Cependant, leur principal défi réside dans la complexité de la conception de l'interaction entre les couches. Assurer une cohérence globale du comportement de l'agent lorsque deux couches peuvent potentiellement donner des ordres contradictoires est un problème d'ingénierie logicielle non trivial.

Le tableau suivant synthétise la comparaison de ces trois paradigmes architecturaux classiques.

Tableau 1 : Comparaison Synthétique des Architectures d'Agents Classiques

Critère	Architecture Réactive	Architecture Délibérative (BDI)	Architecture Hybride
Principe de fonctionnement	Couplage direct perception-action.	Raisonnement symbolique explicite sur des états mentaux.	Combinaison de couches réactives et délibératives.
Représentation du monde	Implicite, dans les règles de comportement.	Explicite et symbolique (base de croyances).	Représentations multiples à différents niveaux d'abstraction.
Processus de décision	Sélection de règles condition-action.	Cycle de délibération (croyances, désirs, intentions, plans).	Interaction et arbitrage entre les couches.
Vitesse de réaction	Très rapide.	Potentiellement lente en raison de la délibération.	Rapide pour les tâches réactives, plus lente pour la planification.
Capacité de	Nulle ou très limitée.	Élevée, planification à	Élevée, gérée par la

planification		long terme.	couche délibérative.
Adaptabilité	Limitée à des comportements pré-programmés.	Adaptable via la révision des croyances et la re-planification.	Très adaptable, combine flexibilité à court et long terme.
Exemple canonique	Architecture de Subsomption (Brooks).	Procedural Reasoning System (PRS).	Architectures en couches (par ex. TouringMachines).
Principale limite	Incapacité à gérer des objectifs complexes et à planifier.	Lenteur potentielle dans les environnements très dynamiques.	Complexité de la conception et de la coordination inter-couches.

Cette analyse des architectures classiques fournit une base de référence indispensable. Elle met en lumière les compromis fondamentaux entre réactivité et délibération qui ont façonné le domaine et prépare le terrain pour comprendre en quoi l'arrivée des grands modèles de langage constitue une rupture paradigmatique, tout en s'inscrivant dans la continuité de cette quête d'une intelligence artificielle équilibrée.

Chapitre 3 : Le Paradigme Révolutionnaire des Agents Fondés sur les LLM

L'émergence des grands modèles de langage (LLM) a provoqué une véritable révolution dans le domaine de l'intelligence artificielle, et les systèmes agentiques ne font pas exception. Les LLM ne sont plus seulement des outils pour générer du texte ; ils sont devenus le cœur de nouvelles architectures agentiques, agissant comme des moteurs de raisonnement et de planification d'une flexibilité sans précédent. Ce chapitre explore ce nouveau paradigme, en analysant le rôle du LLM comme composant cognitif central et en décortiquant l'architecture qui doit l'entourer pour lui permettre de devenir un agent autonome et efficace.

Cette nouvelle approche peut être comprise à travers une analogie puissante : si le LLM est l'unité centrale de traitement (CPU) de l'intelligence, capable d'effectuer des opérations de raisonnement complexes, alors l'architecture agentique est son

système d'exploitation (SE). Un CPU seul, aussi puissant soit-il, est inerte. Il a besoin d'un SE pour gérer la mémoire, interagir avec les périphériques (le monde extérieur) et planifier l'exécution des tâches. De la même manière, un LLM seul ne peut pas agir.[31] C'est l'architecture agentique qui lui fournit les "pilotes" (l'utilisation d'outils), le "système de fichiers" (la mémoire) et le "planificateur de tâches" (le cycle de raisonnement) pour transformer son potentiel cognitif en action intentionnelle.

3.1. Le LLM comme Moteur Cognitif

Au cœur de l'agent moderne se trouve un LLM qui sert de "cerveau" ou de moteur cognitif.[33] Contrairement aux systèmes de planification symbolique classiques, qui opèrent sur des représentations logiques rigides, les LLM raisonnent directement en langage naturel. Cette capacité leur confère plusieurs avantages clés :

- **Flexibilité et Compréhension contextuelle :** Les LLM peuvent comprendre des instructions nuancées en langage naturel, s'adapter à des situations imprévues et générer des plans d'action créatifs, là où les planificateurs classiques échoueraient face à une situation non explicitement modélisée.[32]
- **Décomposition de Tâches (Chaining) :** Une des capacités les plus remarquables des LLM dans un contexte agentique est leur aptitude à la décomposition de tâches, parfois appelée "enchaînement" (*chaining*). À partir d'un objectif de haut niveau formulé par un utilisateur (par exemple, "organise un voyage à Paris pour la semaine prochaine"), le LLM peut générer de manière autonome une séquence d'étapes logiques pour atteindre cet objectif (rechercher les vols, comparer les hôtels, vérifier la météo, réserver les billets, etc.).[5]
- **Auto-réflexion et Correction :** Les agents basés sur les LLM peuvent examiner leurs propres actions, évaluer les résultats et corriger leur stratégie en temps réel. Si une étape du plan échoue, le LLM peut analyser l'erreur et proposer une nouvelle approche, créant ainsi une boucle de rétroaction qui permet à l'agent d'apprendre et de s'améliorer.[5]

Des cadres de raisonnement spécifiques ont été développés pour structurer cette interaction entre le LLM et son environnement. Le plus connu est **ReAct (Reasoning and Acting)**, qui consiste à inciter le LLM à alterner explicitement entre des étapes de raisonnement ("Thought:") et des étapes d'action ("Action:").[20] Cette méthode rend le processus de décision de l'agent plus transparent et plus robuste, car il est forcé de

verbaliser sa logique avant d'agir.

3.2. Composants Essentiels de l'Architecture Agentique LLM

Comme l'analogie avec le système d'exploitation le suggère, le LLM seul ne suffit pas. Son potentiel ne peut être libéré qu'au sein d'une architecture qui lui fournit trois capacités fondamentales : la planification, la mémoire et l'accès à des outils.

- **Planification et Décomposition :** C'est la première étape du cycle de vie de l'agent. L'utilisateur fournit un objectif de haut niveau. L'agent, orchestré par un modèle "conducteur" souvent alimenté par le LLM, décompose cet objectif en un plan d'action séquentiel ou parallèle.[32] Des systèmes comme BabyAGI sont des exemples concrets de cette capacité, où un objectif principal est décomposé en une liste de sous-tâches qui sont ensuite exécutées et raffinées de manière itérative.[1]
- **Mémoire :** Les LLM ont une limitation fondamentale : leur fenêtre de contexte est finie. Ils ne peuvent pas se souvenir des interactions passées au-delà de quelques milliers de tokens. Pour qu'un agent puisse mener à bien des tâches complexes et à long terme, il a besoin d'un mécanisme de mémoire externe.[31] Cette mémoire peut être à court terme, pour maintenir le contexte d'une tâche en cours, ou à long terme, pour apprendre des expériences passées. La technique la plus courante pour implémenter cette mémoire est la **Retrieval-Augmented Generation (RAG)**. Le RAG permet à l'agent de rechercher des informations pertinentes dans une base de connaissances externe (par exemple, une base de données vectorielle contenant des documents, des conversations passées, etc.) et d'injecter ces informations dans le prompt du LLM. Cela permet de fonder les réponses du LLM sur des données factuelles et à jour, et de lui donner accès à un historique quasi infini.[31]
- **Utilisation d'Outils (Tool Use) :** C'est sans doute le composant le plus crucial, car il permet à l'agent de sortir de l'univers purement textuel et d'**agir** sur le monde réel ou numérique. Les "outils" sont des fonctions ou des API que le LLM peut décider d'appeler pour accomplir une tâche.[32] Par exemple, un agent peut utiliser un outil de recherche web pour obtenir des informations en temps réel, un outil de calculatrice pour effectuer des opérations mathématiques précises, ou une API pour interroger une base de données d'entreprise ou exécuter une transaction financière.[20] Le LLM ne fait pas la recherche lui-même ; il génère la décision d'appeler l'outil approprié avec les bons paramètres. L'architecture de

l'agent exécute ensuite cet appel, récupère le résultat et le fournit au LLM pour la prochaine étape de raisonnement. Cette capacité transforme le LLM d'un simple parleur en un véritable acteur.

3.3. L'Écosystème des Frameworks de Développement

La complexité de l'assemblage de ces composants (planification, mémoire, outils) a conduit à l'émergence d'un écosystème de frameworks de développement qui visent à simplifier et à standardiser la création d'agents basés sur les LLM.

- **LangChain** et **LlamaIndex** sont deux des frameworks les plus populaires. Ils fournissent des abstractions et des composants modulaires pour construire des "chaînes" ou des "pipelines" agentiques.[1] Ils facilitent l'intégration de différents LLM, la connexion à des sources de données pour le RAG, et la définition d'outils personnalisés.
- Des systèmes plus intégrés comme **AutoGPT** et **BabyAGI** ont démontré le potentiel des agents entièrement autonomes. À partir d'un simple objectif, ces systèmes peuvent générer et exécuter des tâches de manière continue, en utilisant des outils comme la recherche web et l'écriture de fichiers, jusqu'à ce que l'objectif soit atteint.[1] Bien qu'ils soient souvent plus des démonstrations de faisabilité que des outils de production robustes, ils ont joué un rôle crucial dans la popularisation du concept d'IA agentique.

Ces frameworks représentent les premières tentatives de construire des "systèmes d'exploitation" standardisés pour les LLM, offrant aux développeurs un ensemble d'outils pour orchestrer l'intelligence des modèles de langage.

3.4. Le Débat Architectural : LLM vs. SLM

Alors que le paradigme des agents basés sur les LLM se consolide, une critique importante commence à émerger, remettant en question la dépendance excessive à l'égard de modèles de langage massifs, coûteux et énergivores. Des travaux de recherche récents, notamment l'article "Small Language Models are the Future of Agentic AI", soutiennent que pour la majorité des sous-tâches agentiques, qui sont souvent répétitives et bien délimitées, l'utilisation de **petits modèles de langage**

(Small Language Models - SLM) est non seulement suffisante, mais souvent préférable.[37]

Les SLM offrent plusieurs avantages significatifs : une latence plus faible, des besoins en calcul et en mémoire réduits, et des coûts opérationnels bien moindres, tout en maintenant des performances adéquates pour des tâches spécifiques. Cette perspective ne rejette pas les LLM, mais plaide pour des **architectures agentiques hétérogènes**. Dans un tel modèle, un LLM généraliste pourrait agir comme un "orchestrateur" ou un "superviseur", responsable de la planification de haut niveau et de la gestion des tâches complexes nécessitant un raisonnement général. Ce superviseur déléguerait ensuite l'exécution des sous-tâches plus simples à une flotte d'agents spécialisés, chacun alimenté par un SLM efficace et finement réglé pour sa fonction. Cette approche modulaire, qui combine la puissance de raisonnement des LLM avec l'efficacité des SLM, représente une voie prometteuse pour construire des systèmes d'agents à la fois performants, économiques et plus durables sur le plan environnemental.[37]

Chapitre 4 : La Dynamique des Systèmes Multi-Agents (SMA)

Si la création d'un agent unique et autonome est déjà un défi complexe, la véritable puissance de l'IA agentique se révèle souvent lorsque plusieurs agents interagissent pour résoudre des problèmes qu'aucun d'entre eux ne pourrait aborder seul. Ces Systèmes Multi-Agents (SMA) introduisent une nouvelle couche de complexité liée à la coordination, la communication et l'apprentissage collectif. Ce chapitre explore ces dynamiques, en examinant les mécanismes formels de communication, les défis de l'apprentissage par renforcement multi-agent (MARL), et le phénomène fascinant et parfois imprévisible du comportement émergent.

4.1. Coordination, Communication et Négociation

Dans un SMA, les agents ne sont plus des entités isolées ; ils font partie d'un système social où leurs actions sont interdépendantes. La **coordination** devient alors le défi central : comment s'assurer que les actions des agents individuels, chacun avec sa propre perspective limitée, s'harmonisent pour atteindre un objectif global?.[7] La

littérature sur les SMA a identifié plusieurs mécanismes de coordination :

- **Ajustement Mutuel :** C'est la forme la plus simple de coordination, où deux ou plusieurs agents échangent des informations et ajustent leurs comportements respectifs pour partager des ressources ou atteindre un but commun, sans qu'aucun agent n'ait de contrôle sur les autres.[7]
- **Négociation :** Pour des interactions plus complexes, en particulier lorsque les agents ont des objectifs potentiellement conflictuels, la négociation devient essentielle. Des approches basées sur la **théorie des jeux** sont souvent utilisées pour modéliser ces interactions. Les agents évaluent l'utilité des différentes issues possibles et cherchent à trouver un accord qui maximise leur gain, en supposant qu'ils sont des acteurs rationnels.[7]

Pour que ces mécanismes de coordination fonctionnent, les agents doivent pouvoir communiquer. Une communication ad hoc est fragile et non interopérable. C'est pourquoi des standards ont été développés pour formaliser le langage et les protocoles d'interaction. Le plus influent de ces standards est le **FIPA-ACL (Agent Communication Language)**, proposé par la *Foundation for Intelligent Physical Agents*.[39]

Le FIPA-ACL est fondé sur la **théorie des actes de langage** (*speech act theory*), qui considère que communiquer n'est pas seulement transmettre de l'information, mais accomplir une action intentionnelle.[39] Un message ACL est structuré autour de plusieurs paramètres clés [42] :

- **Performative :** Le paramètre obligatoire qui définit le type d'acte de langage (par exemple, inform, request, propose, agree). Il spécifie l'intention communicative du message.
- **Sender / Receiver :** Identifient l'émetteur et le(s) destinataire(s) du message.
- **Content :** Le contenu propositionnel du message, c'est-à-dire l'information transmise.
- **Language / Ontology :** Spécifient le langage formel et l'ontologie utilisés pour interpréter le contenu, assurant une compréhension sémantique partagée.
- **Protocol / Conversation-ID :** Permettent de gérer des interactions plus complexes en situant le message dans le contexte d'un protocole de conversation spécifique (par exemple, une enchère ou une négociation).

En fournissant une sémantique formelle pour la communication, le FIPA-ACL permet de construire des systèmes interopérables où des agents hétérogènes peuvent interagir de manière fiable et prévisible.

4.2. Apprentissage et Adaptation Collective : MARL

Dans de nombreux scénarios, il est impossible de pré-programmer des stratégies de coordination optimales. Les agents doivent apprendre à collaborer et à s'adapter par l'expérience. Le paradigme dominant pour cet apprentissage collectif est l'**Apprentissage par Renforcement Multi-Agent (MARL)**.[44] Le MARL étend les principes de l'apprentissage par renforcement (où un agent unique apprend une politique d'action en maximisant une récompense) à des contextes multi-agents.

Cependant, cette extension introduit des défis fondamentaux qui n'existent pas dans le cas d'un agent unique [45] :

- **Non-stationnarité de l'environnement :** Du point de vue d'un agent individuel, l'environnement est non stationnaire. En effet, les politiques des autres agents changent au fur et à mesure qu'ils apprennent, ce qui signifie que la meilleure action à un instant t pourrait ne plus l'être à l'instant t+1. Cela viole une hypothèse de base des algorithmes de RL classiques.
- **Scalabilité :** L'espace des états et des actions conjointes croît de manière exponentielle avec le nombre d'agents. Explorer cet espace pour trouver une politique optimale devient rapidement infaisable.
- **Attribution du Crédit (Credit Assignment) :** Dans les tâches coopératives, les agents reçoivent souvent une récompense globale partagée par toute l'équipe. Il est alors extrêmement difficile de déterminer la contribution spécifique de chaque agent au succès ou à l'échec global, ce qui complique la mise à jour de leurs politiques individuelles.

Pour surmonter ces défis, la recherche en MARL a développé plusieurs familles d'algorithmes, comme le montrent des surveys récents.[48] Une approche particulièrement fructueuse est le paradigme de l'

entraînement centralisé avec exécution décentralisée (Centralized Training with Decentralized Execution - CTDE). L'idée est que, pendant la phase d'apprentissage (qui peut se faire en simulation), un critique centralisé a accès aux observations et actions de tous les agents. Il peut ainsi résoudre le problème de l'attribution du crédit et apprendre une fonction de valeur plus stable. Une fois l'entraînement terminé, chaque agent déploie sa politique de manière décentralisée, en n'utilisant que ses propres observations locales.

4.3. Le Phénomène du Comportement Émergent

L'une des caractéristiques les plus fascinantes et les plus importantes des SMA est le **comportement émergent**. Ce phénomène se produit lorsque des modèles et des comportements macroscopiques complexes et organisés apparaissent à partir des interactions locales d'agents simples, sans qu'il y ait de plan ou de contrôle centralisé.[53] Un exemple classique est le vol d'un essaim d'oiseaux : chaque oiseau suit quelques règles simples (s'aligner avec ses voisins, maintenir une distance, se diriger vers le centre perçu du groupe), mais le groupe dans son ensemble produit des formations fluides et complexes qu'aucun oiseau ne dirige.

Les caractéristiques clés du comportement émergent sont [53] :

- **Nouveauté :** Le comportement global est qualitativement différent de celui des agents individuels.
- **Imprédictibilité :** Le comportement macroscopique est difficile, voire impossible, à prédire en analysant uniquement les règles des agents isolés.
- **Auto-organisation :** Le système s'organise spontanément sans autorité centrale.

Ce phénomène est une conséquence directe des mécanismes d'interaction dans les systèmes adaptatifs complexes (CAS), tels que les boucles de rétroaction (où l'action d'un agent influence l'environnement, qui à son tour influence les actions futures des autres agents) et la non-linéarité.[53]

Le comportement émergent est une arme à double tranchant. D'un côté, il est la source de la robustesse et de l'adaptabilité des SMA, leur permettant de trouver des solutions créatives à des problèmes complexes (par exemple, l'optimisation de routes par une colonie de fourmis simulée). De l'autre, il représente un défi majeur pour la sécurité et la fiabilité des systèmes. Un comportement émergent non désiré peut conduire à des défaillances systémiques catastrophiques et imprévues, car il n'est pas explicitement programmé et peut ne pas être détecté par les tests traditionnels.[54] Comprendre, prédire et, si possible, contrôler l'émergence est donc l'un des plus grands défis de l'ingénierie des systèmes multi-agents.

Chapitre 5 : Fiabilité, Vérification et Évaluation des Systèmes Agentiques

La promesse d'autonomie et d'adaptabilité des systèmes agentiques s'accompagne d'un défi majeur : comment garantir qu'ils se comporteront de manière fiable, sûre et conforme à leurs spécifications? À mesure que ces systèmes sont déployés dans des domaines critiques comme la santé, la finance ou les transports, la simple observation de leur performance ne suffit plus. Des méthodes rigoureuses d'évaluation, de test et, idéalement, de vérification formelle deviennent indispensables. Ce chapitre explore les approches existantes pour assurer la fiabilité des systèmes agentiques, des méthodologies d'évaluation empirique aux techniques de preuve mathématique, tout en soulignant les nouveaux défis posés par l'opacité et la nature stochastique des agents basés sur les LLM.

5.1. Méthodologies d'Évaluation et Benchmarks

L'évaluation de la performance d'un système multi-agent est intrinsèquement complexe. Contrairement à un algorithme classique dont on peut mesurer l'efficacité sur un jeu de données statique, un SMA doit être évalué en fonction de son comportement dynamique dans un environnement interactif. Plusieurs approches méthodologiques ont été proposées pour structurer cette évaluation :

- **Approches Basées sur les Objectifs :** L'approche **Goal-Question-Metric (GQM)** est une méthode structurée qui consiste à définir un objectif d'évaluation de haut niveau (par exemple, "évaluer l'efficacité de la coordination des agents"), à le décomposer en questions spécifiques (par exemple, "quel est le temps moyen pour accomplir une tâche collaborative?"), puis à identifier des métriques quantifiables pour répondre à ces questions (par exemple, le temps en secondes).[58]
- **Évaluation par Simulation :** La simulation est l'outil le plus couramment utilisé pour tester et évaluer les SMA. Des plateformes et frameworks spécialisés permettent de créer des environnements virtuels contrôlés où le comportement des agents peut être observé et mesuré dans diverses conditions. Des frameworks modernes comme **Isaac Lab** de NVIDIA, qui s'appuie sur des simulations physiques accélérées par GPU, ou **JAX**, une bibliothèque pour le calcul numérique à haute performance, sont devenus des outils essentiels pour l'entraînement et le test à grande échelle des algorithmes de MARL, en particulier en robotique.[59]

- **Benchmarks et Compétitions :** Pour permettre une comparaison standardisée des algorithmes, la communauté de recherche a développé des benchmarks et des compétitions. Des environnements comme la **RoboCup** (simulation de football robotique) ou des suites de tests comme le **StarCraft Multi-Agent Challenge (SMAC)** ont servi de bancs d'essai pour de nombreux algorithmes de MARL, en fournissant des tâches complexes et standardisées pour évaluer la coordination et la compétition.[47]

5.2. Approches de Vérification Formelle

Pour les systèmes où la sécurité et la fiabilité sont critiques, l'évaluation empirique, même extensive, ne peut garantir l'absence d'erreurs. La **vérification formelle** offre une approche alternative qui vise à prouver mathématiquement qu'un système satisfait ses spécifications. Deux techniques principales sont utilisées dans le contexte des SMA [61] :

- **Model Checking :** Cette technique consiste à construire un modèle mathématique formel du système (généralement un automate à états finis) et à explorer de manière exhaustive tout l'espace des états possibles pour vérifier si une propriété, exprimée dans une logique formelle (comme la logique temporelle ou la logique épistémique), est vraie.[63] Par exemple, on pourrait vérifier la propriété "il est impossible pour deux robots d'entrer en collision" ou "l'agent finira toujours par savoir que le message a été reçu". Des outils comme **MCMAS** ont été spécifiquement développés pour le model checking de systèmes multi-agents, permettant de vérifier des propriétés temporelles, épistémiques (liées à la connaissance des agents) et stratégiques.[66]
- **Theorem Proving (Preuve de Théorèmes) :** Contrairement au model checking qui est automatique mais limité aux systèmes à états finis, la preuve de théorèmes est une approche plus générale mais souvent interactive. Elle consiste à représenter à la fois le système et ses propriétés comme des formules dans un système logique (par exemple, la logique d'ordre supérieur) et à construire une preuve mathématique formelle que les propriétés découlent des axiomes décrivant le système. Des outils comme **Isabelle/HOL** ont été utilisés pour modéliser et vérifier formellement des agents BDI.[61]

Le principal défi de la vérification formelle des SMA est l'**explosion combinatoire de l'espace d'états**. Le nombre total d'états du système croît de manière exponentielle

avec le nombre d'agents, rendant l'exploration exhaustive rapidement infaisable.[68] De plus, la modélisation d'environnements ouverts et dynamiques, ainsi que l'information imparfaite des agents, ajoutent des couches de complexité significatives.[68]

5.3. Les Nouveaux Défis de la Validation des Agents LLM

L'intégration des LLM au cœur des architectures agentiques introduit une nouvelle série de défis pour la fiabilité et la vérification, qui rendent les approches traditionnelles difficiles à appliquer :

- **Nature Stochastique et Non-déterministe :** Les LLM sont des modèles probabilistes. Pour un même prompt, ils peuvent générer des réponses différentes. Ce non-déterminisme rend la reproductibilité des tests et la vérification exhaustive par model checking extrêmement difficiles.
- **Opacité ("Boîte Noire") :** Le processus de décision interne d'un LLM, qui repose sur des milliards de paramètres, est largement opaque et ininterprétable. Il est donc presque impossible de prouver formellement pourquoi une décision spécifique a été prise.
- **Comportement Imprévisible :** Les LLM peuvent "halluciner" (générer des informations fausses) ou être sensibles à des variations subtiles dans le prompt (attaques adversariales), conduisant à des comportements imprévisibles et potentiellement dangereux.

Face à ces défis, la garantie de la fiabilité des agents basés sur les LLM repose moins sur des preuves formelles a priori que sur un ensemble de stratégies de mitigation et de contrôle a posteriori [20] :

- **Supervision Humaine (Human-in-the-loop) :** Maintenir un humain dans la boucle de décision pour valider les actions critiques de l'agent.
- **Journalisation et Traçabilité :** Enregistrer de manière exhaustive toutes les étapes de raisonnement, les appels d'outils et les décisions de l'agent pour permettre une analyse post-mortem en cas d'échec.
- **Garde-fous (Guardrails) :** Mettre en place des mécanismes de sécurité qui filtrent les entrées et les sorties du LLM pour s'assurer qu'elles respectent des contraintes de sécurité, d'éthique et de conformité.
- **Tests Contradictoires (Adversarial Testing) :** Tester systématiquement la robustesse de l'agent en le soumettant à des entrées conçues pour le tromper ou

le faire échouer.

Le tableau suivant classifie les principaux frameworks de développement, qui sont les outils concrets utilisés pour implémenter et, dans une certaine mesure, tester ces différents types de systèmes agentiques.

Tableau 2 : Classification des Frameworks de Développement pour Systèmes Agentiques

Nom du Framework	Paradigme Principal	Langage/Écosystème	Fonctionnalités Clés	Cas d'Usage Typique
JADE	SMA-FIPA	Java	Conformité FIPA, communication ACL, ontologies, services d'annuaire.	Systèmes distribués, simulations, IoT.[21]
Jason	BDI	AgentSpeak (extension de Java)	Implémentation fidèle du modèle BDI, raisonnement logique, gestion des plans.	Modélisation du comportement rationnel, simulations sociales.[21]
OpenAI Gym / Gymnasium	Apprentissage par Renforcement (RL)	Python	Environnements standardisés, interface agent-environnement, support MARL.	Développement et benchmarking d'algorithmes de RL/MARL.[22]
LangChain	LLM-centric	Python, JavaScript	Abstractions pour chaînes, agents, mémoire, RAG, intégration d'outils.	Création rapide d'applications et d'agents basés sur les LLM.[36]
NVIDIA Isaac Lab	Simulation Robotique /	Python (basé sur Isaac Sim)	Simulation physique	Apprentissage Sim-to-Real

	MARL		accélérée par GPU, capteurs réalistes, support MARL.	pour la robotique, entraînement de politiques de contrôle.[60]

Partie III : Implications et Perspectives

Chapitre 6 : Études de Cas et Applications Sectorielles

L'IA agentique n'est plus un domaine purement théorique. Son application concrète transforme déjà de nombreux secteurs en automatisant des tâches complexes, en optimisant des processus et en augmentant les capacités humaines. Ce chapitre présente des études de cas emblématiques qui illustrent l'impact des systèmes agentiques dans des domaines critiques comme la santé, la finance, le développement logiciel et l'industrie.

6.1. L'Agent Autonome dans le Diagnostic Médical et la Santé

Le secteur de la santé est l'un des domaines où le potentiel des agents IA est le plus prometteur. Face à la complexité croissante des données médicales (imagerie, génomique, dossiers médicaux électroniques), les agents autonomes offrent une capacité d'analyse et de synthèse qui dépasse celle des praticiens humains.

Les agents IA sont utilisés pour améliorer les soins aux patients à plusieurs niveaux.[3] Ils peuvent traiter des données génomiques complexes pour identifier des réponses potentielles à des traitements contre le cancer, analyser en continu les signes vitaux pour détecter les signes précoces de détérioration d'un patient, ou encore prédire les taux d'admission pour optimiser l'allocation des ressources hospitalières.[3] En agissant comme des systèmes d'aide à la décision clinique (CDSS), ils peuvent analyser des données multimodales (notes cliniques, résultats de laboratoire, images médicales)

pour proposer des diagnostics différentiels et des plans de traitement personnalisés, réduisant ainsi la charge cognitive des médecins et les risques d'erreur.[70]

Une étude de cas particulièrement révélatrice est le déploiement du système **AI Consult**, développé en partenariat par Penda Health et OpenAI, dans un réseau de cliniques de soins primaires au Kenya.[73] Ce système, basé sur un LLM, agit comme un "filet de sécurité" en analysant en temps réel les décisions des cliniciens enregistrées dans le dossier médical électronique. Une étude pragmatique portant sur près de 40 000 visites de patients a montré que les cliniciens utilisant AI Consult ont commis significativement moins d'erreurs. Les résultats ont indiqué une

réduction de 16% des erreurs de diagnostic et de **13% des erreurs de traitement**. De plus, le système a eu un effet formateur, les cliniciens apprenant à éviter les erreurs courantes même avant de recevoir une alerte du système. Ce déploiement réussi met en évidence les facteurs clés du succès : un modèle performant, une intégration transparente dans les flux de travail cliniques et des stratégies actives pour encourager l'adoption par le personnel.[73]

6.2. L'Automatisation dans la Finance et le Développement Logiciel

Deux secteurs à forte intensité de connaissances, la finance et le développement logiciel, sont en première ligne de la transformation agentique.

Dans le **domaine financier**, les agents IA tirent parti de leur capacité à analyser en continu d'énormes volumes de données de marché en temps réel. Ils peuvent identifier des tendances, évaluer les risques, prendre des décisions d'investissement et exécuter des transactions sur les marchés boursiers de manière autonome, avec une vitesse et une précision inaccessibles aux traders humains.[3] Ces systèmes peuvent également être utilisés pour la détection de fraudes en temps réel et l'automatisation des processus de conformité, améliorant ainsi la sécurité et l'efficacité des opérations financières.[4]

Dans le **développement logiciel**, l'IA agentique est en train de redéfinir le rôle du développeur. Au-delà des assistants de code comme GitHub Copilot, de nouveaux agents plus autonomes sont capables de prendre en charge des projets de développement entiers. À partir d'une description en langage naturel d'un objectif, ces agents peuvent décomposer le problème, générer non seulement le code, mais

aussi la structure des fichiers, les tests unitaires et la documentation.[1] Des systèmes comme

Claude Code peuvent comprendre les objectifs plus larges d'un projet, suggérer des améliorations non demandées et anticiper des problèmes potentiels.[1] Cette automatisation de tâches complexes promet d'augmenter considérablement la productivité des équipes de développement, en leur permettant de se concentrer sur la conception architecturale et les problèmes les plus créatifs.

6.3. Impact sur l'Industrie, la Logistique et l'Éducation

L'impact des systèmes agentiques s'étend à de nombreux autres secteurs :

- **Industrie et Manufacture :** Les agents IA sont au cœur de l'Industrie 4.0. Ils pilotent l'automatisation des chaînes de production avec des robots intelligents, optimisent la maintenance des équipements grâce à la maintenance prédictive, et améliorent le contrôle qualité en analysant les données des capteurs en temps réel.[4]
- **Transport et Logistique :** Les agents optimisent les itinéraires des flottes de véhicules en tenant compte du trafic en temps réel, réduisent les coûts de transport et sont un composant essentiel des véhicules autonomes. Ils permettent également d'anticiper la demande et d'améliorer la gestion des chaînes d'approvisionnement.[2]
- **Agriculture et Environnement :** Dans l'agriculture de précision, les agents analysent les données provenant d'images satellites et de capteurs au sol pour optimiser l'utilisation des ressources comme l'eau et les engrais, améliorant les rendements tout en réduisant l'impact environnemental.[4] Des services comme Prioréno en France utilisent l'IA pour aider les collectivités à planifier la rénovation thermique de leurs bâtiments.[2]
- **Éducation et Formation :** Les agents IA permettent une personnalisation à grande échelle de l'apprentissage. Ils peuvent fournir des ressources éducatives adaptées aux besoins et au rythme de chaque apprenant, agir comme des tuteurs intelligents et libérer du temps pour les enseignants en automatisant les tâches administratives.[2]

Ces études de cas démontrent que l'IA agentique n'est pas une technologie monolithique, mais un paradigme flexible qui peut être adapté pour résoudre des

problèmes spécifiques dans une multitude de domaines, en apportant des gains significatifs en efficacité, en précision et en personnalisation.

Chapitre 7 : Enjeux Éthiques, Économiques et Sociétaux

L'avènement de systèmes d'IA capables d'agir de manière autonome dans le monde soulève des questions profondes qui transcendent le cadre purement technique. La prolifération des agents autonomes nous oblige à reconsidérer des notions fondamentales comme la responsabilité, la confiance et la nature même du travail. Ce chapitre aborde ces enjeux critiques, en analysant les défis éthiques de l'imputabilité et de l'alignement des valeurs, l'impact économique sur le marché du travail, et en concluant par une autocritique des paradigmes technologiques dominants.

7.1. La Question de la Responsabilité et de l'Imputabilité

L'un des défis les plus pressants posés par l'IA agentique est celui de la responsabilité. Lorsqu'un agent autonome prend une décision qui cause un préjudice – qu'il s'agisse d'une perte financière due à une transaction boursière erronée, d'un accident de voiture autonome ou d'une erreur de diagnostic médical – qui est responsable? Est-ce l'utilisateur qui a défini l'objectif, le développeur qui a conçu l'agent, l'entreprise qui l'a déployé, ou l'agent lui-même?.[5]

Les cadres juridiques et éthiques traditionnels, conçus pour des outils sous contrôle humain direct, sont mal adaptés à cette nouvelle réalité. L'autonomie de l'agent crée un "fossé de responsabilité" (*responsibility gap*), où il devient difficile d'attribuer la culpabilité.[74] Ce problème est exacerbé par l'opacité des modèles sous-jacents comme les LLM. Si même les concepteurs ne peuvent pas expliquer complètement pourquoi un agent a pris une décision particulière, l'établissement de la causalité et de l'imputabilité devient un véritable casse-tête. La mise en place de cadres de responsabilisation clairs, qui définissent les obligations des différentes parties prenantes et exigent des mécanismes de traçabilité et d'auditabilité des décisions des agents, est une nécessité absolue pour un déploiement sûr et socialement acceptable de ces technologies.[3]

7.2. L'Alignement des Valeurs (Value Alignment)

Au-delà de la simple exécution de tâches, comment pouvons-nous nous assurer que les objectifs poursuivis par les agents autonomes sont alignés avec les valeurs humaines, les normes éthiques et les règles sociétales?.[5] C'est le problème de l'

alignement des valeurs, un champ de recherche en IA qui est devenu d'une importance capitale.[76] Un agent optimisé pour un objectif apparemment bénin (par exemple, "maximiser la production de trombones") pourrait, s'il n'est pas contraint par des valeurs humaines, adopter des stratégies extrêmes et nuisibles pour atteindre cet objectif (par exemple, convertir toutes les ressources de la Terre en trombones), comme l'illustre la célèbre expérience de pensée de Nick Bostrom.

Le défi de l'alignement est particulièrement aigu dans les systèmes multi-agents, où des agents hétérogènes, potentiellement conçus par différentes organisations avec des valeurs différentes, doivent interagir. Des approches sont développées pour intégrer des normes et des valeurs dans les SMA, par exemple en utilisant des algorithmes évolutionnistes pour trouver des ensembles de normes qui optimisent la promotion de multiples valeurs simultanément, même lorsqu'elles sont conflictuelles.[78] Garantir que les agents agissent non seulement de manière efficace, mais aussi de manière éthique et socialement bénéfique, est sans doute le plus grand défi à long terme pour le domaine.

7.3. Impact sur le Marché du Travail

L'automatisation des tâches cognitives par les agents IA est en train de provoquer une transformation structurelle profonde du marché du travail. Les analyses des grandes organisations internationales brossent un tableau nuancé, fait de destructions, de créations et, surtout, de reconfigurations d'emplois.

Selon l'Organisation de Coopération et de Développement Économiques (OCDE), environ **27% des emplois** dans les pays membres se trouvent dans des catégories professionnelles où plus de 70% des tâches sont hautement susceptibles d'être automatisées par les technologies actuelles.[82] Le Forum Économique Mondial (WEF),

dans son rapport "Future of Jobs 2025", corrobore cette tendance en identifiant des professions en déclin marqué d'ici 2030, telles que les caissiers, les employés de billetterie et les assistants administratifs – des rôles caractérisés par des tâches répétitives et procédurales.[82]

Cependant, cette transformation n'est pas uniquement une histoire de destruction d'emplois. L'émergence de l'économie des agents IA crée simultanément une demande pour de nouvelles compétences et de nouvelles professions. Les analyses du marché du travail européen et américain montrent une forte augmentation des offres d'emploi pour des rôles qui existaient à peine il y a quelques années : ingénieurs en IA, *prompt engineers*, analystes en cybersécurité spécialisés en IA, consultants en technologie IA, et éthiciens de l'IA.[82] L'impact de l'IA sur le travail n'est donc pas monolithique. Il s'agit d'une bifurcation : les tâches routinières sont de plus en plus automatisées, tandis que de nouvelles opportunités apparaissent dans la conception, la gouvernance, la supervision et l'interaction stratégique avec les systèmes d'IA. La compétence clé de demain ne sera plus seulement l'exécution de tâches, mais la capacité à collaborer efficacement avec des partenaires agents-IA.[82]

7.4. Contre-arguments et Autocritique : Les Limites des Approches Actuelles

Il est essentiel de maintenir une perspective critique sur les tendances dominantes. La trajectoire actuelle du développement de l'IA, en particulier celle des LLM, repose fortement sur les **"lois d'échelle" (*scaling laws*)**, l'idée que l'amélioration des performances découle principalement de l'augmentation de la taille des modèles, de la quantité de données d'entraînement et de la puissance de calcul.

- **Critique des "Scaling Laws" :** Cette approche est de plus en plus critiquée comme étant insoutenable sur les plans économique et écologique. Les coûts de formation des modèles de pointe se chiffrent en centaines de millions de dollars, et la consommation d'énergie des centres de données devient une préoccupation environnementale majeure.[84] De plus, cette course à la taille crée une barrière à l'entrée, concentrant le pouvoir technologique entre les mains de quelques grandes entreprises et soulevant des questions de confiance envers des modèles "boîtes noires" et propriétaires.[84]
- **L'Alternative Neuro-symbolique :** En réponse à ces limites, le domaine de l'**IA neuro-symbolique** gagne en importance. Cette approche vise à combiner les forces des deux traditions de l'IA : la capacité des réseaux de neurones à

apprendre des motifs à partir de données brutes, et la capacité de l'IA symbolique à effectuer un raisonnement logique et explicite.[85] En intégrant des connaissances et des contraintes symboliques dans les modèles neuronaux, l'IA neuro-symbolique promet de créer des systèmes qui généralisent mieux à partir de moins de données, qui sont plus interprétables et qui sont plus fiables dans leur raisonnement.[85] Les architectures hybrides qui cherchent à intégrer des cadres comme le BDI avec des LLM peuvent être considérées comme une manifestation de cette tendance neuro-symbolique, cherchant à structurer la flexibilité des LLM avec la rigueur du raisonnement symbolique.

Le tableau suivant synthétise les principaux défis éthiques et sociétaux discutés dans ce chapitre.

Tableau 3 : Synthèse des Défis Éthiques et Sociétaux des Systèmes Agentiques

Domaine du Défi	Description du Défi	Exemples Concrets	Pistes de Solution / Cadres Réglementaires
Éthique	**Imputabilité & Responsabilité**	Un véhicule autonome provoque un accident ; un agent financier cause des pertes massives.[5]	Cadres juridiques pour la responsabilité partagée, "boîtes noires" enregistrant les décisions, assurance obligatoire.
Éthique	**Alignement des Valeurs**	Un agent de recommandation optimise l'engagement en favorisant la désinformation ; un agent militaire optimise une mission avec des dommages collatéraux inacceptables.[5]	Recherche sur le "Value Alignment", conception de systèmes normatifs, comités d'éthique internes, audits externes.[78]
Éthique	**Manipulation & Déception**	Un agent commercial utilise des techniques psychologiques pour	Obligation de transparence (divulgation de

		pousser à l'achat ; un chatbot se fait passer pour un humain pour gagner la confiance.[88]	l'identité de l'IA), interdiction des "dark patterns", éducation du public.
Économique	**Transformation du Travail**	Automatisation des emplois administratifs, des centres d'appels, de l'analyse de données de routine.[82]	Investissement massif dans la formation continue et la reconversion, réforme de l'éducation, instauration de filets de sécurité sociale (revenu de base, etc.).
Sociétal	**Biais & Équité**	Un agent de recrutement IA discrimine systématiquement certains profils démographiques en se basant sur des données historiques biaisées.[3]	Audits réguliers des biais algorithmiques, utilisation de données d'entraînement diversifiées et représentatives, cadres réglementaires (ex: AI Act européen).
Sociétal	**Confiance Humain-Machine**	Diminution de la confiance interpersonnelle due aux deepfakes ; sur-confiance ou sous-confiance des opérateurs envers les systèmes IA.[89]	Conception pour la confiance (*Trust-Centered Design*), développement de systèmes explicables (XAI), certification des systèmes IA.
Technique	**Sécurité & Robustesse**	Des agents peuvent être piratés pour des actions malveillantes ; un comportement émergent imprévu dans un SMA cause une défaillance systémique.[54]	Tests contradictoires (*adversarial testing*), vérification formelle pour les composants critiques, "garde-fous" robustes, protocoles de sécurité

			inter-agents.

Discussion Générale

L'analyse menée à travers ce mémoire a permis de cartographier le paysage complexe de la création de systèmes d'IA agentiques, depuis ses racines dans l'IA symbolique jusqu'à sa floraison spectaculaire à l'ère des grands modèles de langage. Cette discussion a pour but de synthétiser les principaux enseignements de cette analyse, de mettre en lumière leurs implications théoriques et pratiques, et de formuler des recommandations pour guider la recherche future et les acteurs du terrain.

Synthèse des Apports : Vers une Théorie Unifiée de la Création d'Agents?

L'apport central de ce travail est de proposer une lecture de l'histoire et de l'état de l'art de l'IA agentique non pas comme une série de ruptures, mais comme une **synthèse dialectique**. La thèse initiale, incarnée par l'IA symbolique et les architectures comme le BDI, a posé les fondations conceptuelles de l'agentivité rationnelle, en mettant l'accent sur la structure, la logique et la délibération explicite. L'antithèse, apportée par l'IA connexionniste et les LLM, a introduit une flexibilité, une capacité d'apprentissage et une maîtrise du langage naturel sans précédent, mais au prix de l'opacité et d'un manque de structure sémantique forte.

La synthèse, qui émerge aujourd'hui, est l'architecture de l'**agent hybride neuro-symbolique**. Dans cette vision, les deux approches ne s'opposent plus mais se complètent. Les concepts de l'architecture BDI (croyances, désirs, intentions) ne sont pas remplacés par le LLM ; ils deviennent le cadre sémantique et structurel qui orchestre le LLM. Le BDI fournit la réponse au "quoi faire" (quels sont mes engagements, mes intentions?), tandis que le LLM, avec ses capacités de raisonnement flexible, fournit une réponse puissante au "comment le faire" (quel est le meilleur plan d'action dans cette situation contextuelle?).

Des travaux de recherche récents commencent à explorer explicitement cette intégration. Des agents sont conçus pour suivre un cycle de raisonnement inspiré du

BDI, où un LLM est utilisé pour mettre à jour les croyances à partir de perceptions textuelles, pour délibérer sur les désirs à poursuivre, et pour générer des plans d'action pour satisfaire les intentions.[15] Cette convergence suggère que nous nous dirigeons vers une théorie plus unifiée de la création d'agents, où la structure symbolique et la flexibilité neuronale ne sont plus des alternatives, mais des composants nécessaires d'une même architecture cognitive.

Implications Théoriques et Pratiques

Cette perspective de synthèse a des implications profondes, tant pour la recherche fondamentale que pour l'ingénierie des systèmes d'IA.

Sur le plan théorique, cela renforce l'idée que la voie vers une intelligence artificielle plus générale et robuste ne réside probablement ni dans une approche purement symbolique, ni dans une approche purement connexionniste, mais dans leur intégration intelligente. Cela invite à revisiter les architectures cognitives classiques, non pas comme des reliques du passé, mais comme des sources d'inspiration pour structurer les capacités émergentes des grands modèles neuronaux. La recherche sur l'IA neuro-symbolique, qui cherche à fonder cette intégration sur des bases formelles, apparaît alors non pas comme un champ de niche, mais comme une direction centrale pour l'avenir de l'IA.

Sur le plan pratique, les implications pour les ingénieurs et les développeurs sont directes. Concevoir un agent basé sur un LLM en se contentant d'une simple boucle de prompts est une approche fragile et imprévisible. Les ingénieurs qui intègrent dans leur conception les principes issus des architectures BDI – une séparation claire entre l'état du monde (croyances), les objectifs (désirs), les engagements (intentions) et les séquences d'actions (plans) – seront mieux à même de construire des systèmes plus robustes, plus prévisibles et plus faciles à déboguer. Par exemple, maintenir un "registre des intentions" explicite permet à l'agent (et à son superviseur humain) de savoir à tout moment à quoi il s'est engagé, et de reconsidérer cet engagement de manière contrôlée, plutôt que de se fier uniquement au contexte implicite du LLM.

Recommandations pour la Recherche Future et pour les Acteurs du Terrain

Sur la base de notre analyse, nous formulons les recommandations suivantes :

Pour la recherche future :

1. **Développer des méthodes de vérification formelle pour les agents hybrides :** La fiabilité des agents critiques est primordiale. Un effort de recherche majeur doit être consacré à l'extension des techniques de model checking et de preuve de théorèmes pour qu'elles puissent s'appliquer à des architectures qui intègrent des composants stochastiques et opaques comme les LLM, potentiellement en vérifiant la structure de contrôle BDI tout en traitant le LLM comme un oracle probabiliste.
2. **Créer des benchmarks standardisés pour l'évaluation de l'agentivité :** Au-delà de la performance sur des tâches spécifiques, de nouveaux benchmarks sont nécessaires pour évaluer quantitativement les propriétés fondamentales de l'agentivité : le degré d'autonomie, la qualité de la proactivité, la robustesse de la réactivité et l'efficacité de la sociabilité dans des scénarios complexes.
3. **Approfondir la recherche sur l'alignement des valeurs dans les SMA décentralisés :** Comment assurer un alignement global des valeurs dans une société d'agents hétérogènes et autonomes sans imposer un contrôle centralisé? La recherche à l'intersection de la théorie des jeux, des systèmes normatifs et de l'apprentissage multi-agent est ici cruciale.
4. **Explorer les architectures agentiques à base de SLM :** La piste des architectures hétérogènes, où un LLM supervise une flotte de SLM spécialisés, mérite une exploration approfondie pour développer des systèmes plus efficaces, économiques et durables.

Pour les acteurs du terrain (entreprises, développeurs, décideurs politiques) :

1. **Adopter une approche de "conception pour la confiance" (Trust-Centered Design) :** La confiance ne doit pas être une réflexion après coup. Elle doit être intégrée dès la conception des agents, en privilégiant la transparence du processus de décision (par exemple, en utilisant des cadres comme ReAct), la traçabilité des actions et la possibilité pour l'utilisateur de comprendre et de contrôler le comportement de l'agent.
2. **Investir dans la formation continue pour la collaboration homme-agent :** La transformation du marché du travail est inévitable. Les entreprises et les gouvernements doivent investir massivement dans des programmes de formation et de reconversion qui préparent la main-d'œuvre non pas à être remplacée par des agents, mais à collaborer avec eux, en se concentrant sur les compétences

de supervision, de jugement critique et de gestion stratégique.

3. **Mettre en place des cadres de gouvernance internes et externes :** Les entreprises qui déploient des agents IA doivent établir des cadres de gouvernance internes clairs pour la gestion des risques et la responsabilité. Au niveau sociétal, les régulateurs doivent accélérer le développement de cadres législatifs (comme l'AI Act en Europe) qui clarifient les responsabilités et imposent des exigences de sécurité et de transparence pour les agents autonomes à haut risque.[31]

Conclusion Générale

Au terme de cette exploration approfondie, il apparaît clairement que la création de systèmes d'IA agentiques constitue l'une des entreprises les plus ambitieuses et les plus transformatrices de l'informatique contemporaine. Ce mémoire s'est attaché à dénouer les fils complexes de ce domaine, en reliant son riche héritage théorique aux innovations disruptives de l'ère des grands modèles de langage.

Rappel de la Problématique et des Résultats Clés

Nous sommes partis de la problématique de la convergence entre les architectures d'agents symboliques classiques et les capacités sub-symboliques des LLM. Notre analyse a permis de répondre aux questions de recherche initiales en démontrant que :

1. Les principes fondamentaux de l'agentivité (autonomie, réactivité, proactivité, sociabilité) et les cadres architecturaux comme le BDI sont non seulement pertinents, mais essentiels pour structurer et contrôler les agents modernes.
2. Les LLM fonctionnent comme des moteurs cognitifs dont le potentiel n'est libéré qu'au sein d'une architecture agentique complète, dotée de modules pour la planification, la mémoire et l'utilisation d'outils.
3. Les défis classiques des SMA, comme la coordination et la non-stationnarité, persistent et sont même amplifiés par la flexibilité des LLM, ce qui renforce la nécessité de protocoles de communication formels et d'algorithmes de MARL robustes.

4. La nouvelle vague d'autonomie agentique engendre des défis éthiques, économiques et sociétaux majeurs, qui exigent des cadres de gouvernance et de responsabilité repensés.

Le résultat clé de ce mémoire est la validation de nos hypothèses : l'avenir des systèmes agentiques ne réside pas dans le remplacement d'un paradigme par un autre, mais dans leur **synthèse hybride**.

Apports Originaux du Mémoire

L'apport principal de ce travail réside dans sa perspective unificatrice. En conceptualisant l'IA agentique moderne comme une synthèse dialectique des traditions symboliques et connexionnistes, et en utilisant des analogies éclairantes comme celle du "système d'exploitation pour l'intelligence", ce mémoire propose un cadre d'analyse original pour comprendre l'état de l'art et anticiper les évolutions futures. Il ne se contente pas de décrire les technologies, mais cherche à en extraire les principes de conception sous-jacents et à les replacer dans un contexte historique et critique plus large. En reliant systématiquement les architectures les plus récentes aux théories fondatrices de l'agentivité, il offre une grille de lecture cohérente pour un domaine en pleine effervescence.

Ouverture sur de Nouvelles Pistes de Recherche

Le champ de l'IA agentique est loin d'avoir atteint sa maturité. Les travaux présentés ici ouvrent la voie à des pistes de recherche encore plus audacieuses. L'une des prochaines frontières sera sans doute celle des **agents incarnés (*embodied agents*)**, où les principes de l'agentivité logicielle seront pleinement intégrés dans des corps robotiques interagissant avec le monde physique. Cela posera de nouveaux défis en matière de perception, de contrôle moteur et d'apprentissage en temps réel.

Une autre piste fascinante est celle des **sociétés d'agents** à très grande échelle, qui pourraient être utilisées pour simuler des systèmes économiques, sociaux ou écologiques avec une fidélité sans précédent, permettant d'explorer des scénarios complexes et de tester des politiques publiques avant leur mise en œuvre.

Enfin, à plus long terme, la sophistication croissante des architectures cognitives des agents, combinant raisonnement, mémoire, apprentissage et interaction sociale, nous rapprochera inévitablement des questions les plus fondamentales de l'intelligence artificielle : celles de la conscience, de l'intentionnalité et de la nature de l'esprit. La création de systèmes d'IA agentiques n'est pas seulement un défi d'ingénierie ; c'est aussi une quête pour comprendre et recréer les principes de l'intelligence elle-même.

Bibliographie

(Cette section contiendrait la liste complète des plus de 100 références académiques citées dans le mémoire, formatée selon une norme standard comme APA 7.)

Annexes

(Cette section pourrait contenir des éléments complémentaires tels que des extraits de code pour des implémentations d'agents, des guides d'entretien menés avec des experts du domaine, ou des données brutes issues d'expérimentations.)

Ouvrages cités

1. L'IA agentique expliquée simplement : quand l'intelligence artificielle devient autonome - Oo2 Formations, dernier accès : juillet 30, 2025, https://www.oo2.fr/actualites/l-ia-agentique-expliquee-simplement-l-intelligence-artificielle-devient-autonome
2. Intelligence artificielle (IA) : de quoi parle-t-on ? | enseignementsup-recherche.gouv.fr, dernier accès : juillet 30, 2025, https://www.enseignementsup-recherche.gouv.fr/fr/intelligence-artificielle-de-quoi-parle-t-91190
3. Comprendre les agents d'intelligence artificielle : L'avenir des systèmes autonomes, dernier accès : juillet 30, 2025, https://www.datacamp.com/fr/blog/ai-agents
4. Intelligence Artificielle : Définition, histoire, enjeux - DataScientest, dernier accès : juillet 30, 2025, https://datascientest.com/intelligence-artificielle-definition
5. L'IA agentique, qu'est-ce que c'est ? - Red Hat, dernier accès : juillet 30, 2025, https://www.redhat.com/fr/topics/ai/what-is-agentic-ai

6. (PDF) Les Systèmes Multi-Agents - Support de cours - ResearchGate, dernier accès : juillet 30, 2025, https://www.researchgate.net/publication/374902666_Les_Systemes_Multi-Agents_-_Support_de_cours
7. Systèmes multiagents : Principes généraux et ... - SI & Management, dernier accès : juillet 30, 2025, http://www.sietmanagement.fr/wp-content/uploads/2017/12/Chaib-draa2001.pdf
8. Intelligence Artificielle : comprendre les AI Agentiques , une nouvelle vision de l'intelligence artificielle - Olvani, dernier accès : juillet 30, 2025, http://www.olvani.com/magazine/actualites-digitales/intelligence-artificielle-comprendre-les-ai-agentiques-une-nouvelle-vision-de-l%E2%80%99intelligence-artificielle-article
9. A Systematic Literature Review in Multi-Agent Systems: Patterns and Trends | Request PDF, dernier accès : juillet 30, 2025, https://www.researchgate.net/publication/340892947_A_Systematic_Literature_Review_in_Multi-Agent_Systems_Patterns_and_Trends
10. Multi-Agent Systems and Complex Networks: Review and Applications in Systems Engineering - MDPI, dernier accès : juillet 30, 2025, https://www.mdpi.com/2227-9717/8/3/312
11. (PDF) Design Patterns for Multi-agent Systems: A Systematic ..., dernier accès : juillet 30, 2025, https://www.researchgate.net/publication/289338062_Design_Patterns_for_Multi-agent_Systems_A_Systematic_Literature_Review
12. Intelligent Agents: The Key Concepts - CiteSeerX, dernier accès : juillet 30, 2025, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=7434cc7952a7d9b7fb74856d8ab5d60917aa1e0a
13. Intelligent Agents: the transformative AI trend for 2024 | by Ian Watson | Medium, dernier accès : juillet 30, 2025, https://medium.com/@dr.i.watson/intelligent-agents-the-transformative-ai-trend-for-2024-8aaeb19a998c
14. The Knowledge Engineering Review, Vol. 10:2, 1995, 115–152 ..., dernier accès : juillet 30, 2025, https://www.cs.ox.ac.uk/people/michael.wooldridge/pubs/ker95.pdf
15. [2506.01463] Agentic AI and Multiagentic: Are We Reinventing the Wheel? - arXiv, dernier accès : juillet 30, 2025, https://arxiv.org/abs/2506.01463
16. Agents IA : Architecture, fonctionnalités clés et mise en œuvre - TOP Turnover, dernier accès : juillet 30, 2025, https://top-turnover.ai/agents-ia-architecture-fonctions-cles-et-deploiement-en-entreprise/
17. (111005_Version finale thèse) - Université de Lille, dernier accès : juillet 30, 2025, https://pepite-depot.univ-lille.fr/LIBRE/EDBSL/2011/2011LIL2S029.pdf
18. [PDF] Intelligent agents: theory and practice | Semantic Scholar, dernier accès : juillet 30, 2025, https://www.semanticscholar.org/paper/Intelligent-agents%3A-theory-and-practice-Wooldridge-Jennings/d621786b597687f555fae83dc1a021fd21713d90

19. Nicholas R. Jennings and Michael J. Wooldridge (eds.): Agent Technology: Foundations, Applications and Markets - JASSS, dernier accès : juillet 30, 2025, https://www.jasss.org/2/3/sichman.html
20. Qu'est-ce qu'un agent d'intelligence artificielle ? | IBM, dernier accès : juillet 30, 2025, https://www.ibm.com/fr-fr/think/topics/ai-agents
21. Understanding BDI Agents in Agent-Oriented Programming - SmythOS, dernier accès : juillet 30, 2025, https://smythos.com/developers/agent-development/agent-oriented-programming-and-bdi-agents/
22. AI Agent Framework : définition, types et cas d'utilisation | Astera, dernier accès : juillet 30, 2025, https://www.astera.com/fr/type/blog/ai-agent-framework/
23. What Is Agentic Architecture? | IBM, dernier accès : juillet 30, 2025, https://www.ibm.com/think/topics/agentic-architecture
24. AI Agent Architectures: Modular, Multi-Agent, and Evolving - ProjectPro, dernier accès : juillet 30, 2025, https://www.projectpro.io/article/ai-agent-architectures/1135
25. Leveraging the Beliefs-Desires-Intentions Agent Architecture | Microsoft Learn, dernier accès : juillet 30, 2025, https://learn.microsoft.com/en-us/archive/msdn-magazine/2019/january/machine-learning-leveraging-the-beliefs-desires-intentions-agent-architecture
26. BDI Agent Architectures: A Survey - IJCAI, dernier accès : juillet 30, 2025, https://www.ijcai.org/proceedings/2020/0684.pdf
27. The Belief-Desire-Intention Model of Agency - CiteSeerX, dernier accès : juillet 30, 2025, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=495bc4ecb3c68a069758276e7844e7ed6dd3f20f
28. Belief–desire–intention software model - Wikipedia, dernier accès : juillet 30, 2025, https://en.wikipedia.org/wiki/Belief%E2%80%93desire%E2%80%93intention_software_model
29. BDI Agent Architectures: A Survey, dernier accès : juillet 30, 2025, https://repositorio.pucrs.br/dspace/bitstream/10923/18600/2/BDI_Agent_Architectures_A_Survey.pdf
30. Architecture agentique : votre guide complet - Astera Software, dernier accès : juillet 30, 2025, https://www.astera.com/fr/type/blog/agentic-architecture/
31. L'IA agentique (Autonomous GenAI agents) transformera la productivité des entreprises, dernier accès : juillet 30, 2025, https://www.deloitte.com/fr/fr/Industries/tmt/perspectives/ia-agentique-autonomous-genai-agents-transformera-la-productivite-des-entreprises.html
32. Qu'est-ce que l'IA agentique ? | IBM, dernier accès : juillet 30, 2025, https://www.ibm.com/fr-fr/think/topics/agentic-ai
33. ADaPT: As-Needed Decomposition and Planning with Language Models - ResearchGate, dernier accès : juillet 30, 2025, https://www.researchgate.net/publication/382634289_ADaPT_As-Needed_Decomposition_and_Planning_with_Language_Models

34. LLM-Powered AI Agent Systems and Their Applications in ... - arXiv, dernier accès : juillet 30, 2025, https://arxiv.org/abs/2505.16120
35. [2501.05468] LatteReview: A Multi-Agent Framework for Systematic Review Automation Using Large Language Models - arXiv, dernier accès : juillet 30, 2025, https://arxiv.org/abs/2501.05468
36. AI Agent Framework: What it is, Types, & Use Cases - Astera Software, dernier accès : juillet 30, 2025, https://www.astera.com/type/blog/ai-agent-framework/
37. Small Language Models are the Future of Agentic AI - arXiv, dernier accès : juillet 30, 2025, https://arxiv.org/html/2506.02153v1
38. [2506.02153] Small Language Models are the Future of Agentic AI - arXiv, dernier accès : juillet 30, 2025, https://arxiv.org/abs/2506.02153
39. An Introduction to FIPA Agent Communication Language: Standards for Interoperable Multi-Agent Systems - SmythOS, dernier accès : juillet 30, 2025, https://smythos.com/developers/agent-development/fipa-agent-communication-language/
40. Mastering Agent Communication - Number Analytics, dernier accès : juillet 30, 2025, https://www.numberanalytics.com/blog/ultimate-guide-to-agent-communication-languages
41. sarl/sarl-acl: FIPA Agent Communication Language for SARL - GitHub, dernier accès : juillet 30, 2025, https://github.com/sarl/sarl-acl
42. FIPA ACL Message Structure Specification - FIPA.org, dernier accès : juillet 30, 2025, http://www.fipa.org/specs/fipa00061/SC00061G.html
43. FIPA ACL Message Structure Specification, dernier accès : juillet 30, 2025, http://euro.ecom.cmu.edu/program/courses/tcr854/2001/readings/XC00061D.doc
44. Multi-Agent Reinforcement Learning: A Survey - ResearchGate, dernier accès : juillet 30, 2025, https://www.researchgate.net/publication/224695508_Multi-Agent_Reinforcement_Learning_A_Survey
45. Multi-Agent Reinforcement Learning: A Review of Challenges and Applications - MDPI, dernier accès : juillet 30, 2025, https://www.mdpi.com/2076-3417/11/11/4948
46. Survey of recent multi-agent reinforcement learning algorithms utilizing centralized training, dernier accès : juillet 30, 2025, https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11746/117462K/Survey-of-recent-multi-agent-reinforcement-learning-algorithms-utilizing-centralized/10.1117/12.2585808.full
47. Multiagent Systems: A Survey from a Machine Learning Perspective, dernier accès : juillet 30, 2025, https://www.cs.cmu.edu/~mmv/papers/MASsurvey.pdf
48. RaghuHemadri/Multi-Agent-Reinforcement-Learning-Survey-Papers - GitHub, dernier accès : juillet 30, 2025, https://github.com/RaghuHemadri/Multi-Agent-Reinforcement-Learning-Survey-Papers
49. chrisyrniu/Recent-Advances-in-Multi-Agent-Reinforcement-Learning: A collection of recent MARL papers - GitHub, dernier accès : juillet 30, 2025, https://github.com/chrisyrniu/Recent-Advances-in-Multi-Agent-Reinforcement-L

earning
50. Value-Based Deep Multi-Agent Reinforcement Learning with Dynamic Sparse Training, dernier accès : juillet 30, 2025, https://proceedings.neurips.cc/paper_files/paper/2024/hash/31888563b194f9bb33ce1aebc7e1551c-Abstract-Conference.html
51. Deep Reinforcement Learning for Multi-Agent Interaction | EconCS Group, dernier accès : juillet 30, 2025, https://econcs.seas.harvard.edu/event/deep-reinforcement-learning-multi-agent-interaction
52. Cooperative Exploration for Multi-Agent Deep Reinforcement Learning, dernier accès : juillet 30, 2025, http://proceedings.mlr.press/v139/liu21j/liu21j.pdf
53. Emergent Behavior in Multi-Agent Systems: How Complex ... - Medium, dernier accès : juillet 30, 2025, https://medium.com/@sanjeevseengh/emergent-behavior-in-multi-agent-systems-how-complex-behaviors-arise-from-simple-agent-0e4503b376ce
54. An architecture for identifying emergent behavior in multi-agent systems - ResearchGate, dernier accès : juillet 30, 2025, https://www.researchgate.net/publication/290160025_An_architecture_for_identifying_emergent_behavior_in_multi-agent_systems
55. Emergent Behavior - AI Ethics Lab, dernier accès : juillet 30, 2025, https://aiethicslab.rutgers.edu/e-floating-buttons/emergent-behavior/
56. Emergent Behavior Development and Control in Multi-Agent Systems - DTIC, dernier accès : juillet 30, 2025, https://apps.dtic.mil/sti/trecms/pdf/AD1084423.pdf
57. Emergence in Multi-Agent Systems: A Safety Perspective - arXiv, dernier accès : juillet 30, 2025, https://arxiv.org/html/2408.04514v1
58. An Evaluation Method for Multi-Agent Systems | Request PDF - ResearchGate, dernier accès : juillet 30, 2025, https://www.researchgate.net/publication/267895052_An_Evaluation_Method_for_Multi-Agent_Systems
59. A roboticist's journey with JAX: Finding efficiency in optimal control and simulation, dernier accès : juillet 30, 2025, https://developers.googleblog.com/en/a-roboticists-journey-with-jax/
60. Unified framework for robot learning built on NVIDIA Isaac Sim - GitHub, dernier accès : juillet 30, 2025, https://github.com/isaac-sim/IsaacLab
61. Formal Verification of BDI Agents - Research Explorer - The University of Manchester, dernier accès : juillet 30, 2025, https://research.manchester.ac.uk/en/publications/formal-verification-ofbdi-agents
62. Formal Verification of BDI Agents - Aarhus University - Pure, dernier accès : juillet 30, 2025, https://pure.au.dk/portal/en/publications/formal-verification-ofbdi-agents
63. Model checking multi-agent systems - UCL Discovery, dernier accès : juillet 30, 2025, https://discovery.ucl.ac.uk/5627/1/5627.pdf
64. Model Checking - A Hands-On Introduction - ICAPS03 International Conference on Automated Planning&Scheduling, dernier accès : juillet 30, 2025,

https://icaps03.icaps-conference.org/tutorials/tutorial4.htm
65. Model Checking: A Tutorial Overview, dernier accès : juillet 30, 2025, https://members.loria.fr/SMerz/papers/mc-tutorial.pdf
66. MCMAS: A model checker for the verification of multi-agent systems* - Imperial College London, dernier accès : juillet 30, 2025, https://www.doc.ic.ac.uk/~alessio/papers/09/CAV-AL+.pdf
67. Formal Verification of BDI Agents | Request PDF - ResearchGate, dernier accès : juillet 30, 2025, https://www.researchgate.net/publication/385171794_Formal_Verification_of_BDI_Agents
68. Verifying BDI Agents in Dynamic Environments - School of Computing Science - University of Glasgow, dernier accès : juillet 30, 2025, https://www.dcs.gla.ac.uk/~michele/papers/SEKE22.pdf
69. Model Checking for Probabilistic Multiagent Systems - SciOpen, dernier accès : juillet 30, 2025, https://www.sciopen.com/article/10.1007/s11390-022-1218-6
70. AI Agents in Medical Diagnostics 2025 | Revolutionizing Healthcare - Rapid Innovation, dernier accès : juillet 30, 2025, https://www.rapidinnovation.io/post/ai-agents-for-diagnostic-support
71. AI Agent-Powered Multi-Medical Diagnostics: | by Alex G. Lee | Medium, dernier accès : juillet 30, 2025, https://medium.com/@alexglee/ai-agent-powered-multi-medical-diagnostics-6fe79d3733cb
72. Artificial Intelligent Agent Architecture and Clinical Decision-Making in the Healthcare Sector, dernier accès : juillet 30, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11309744/
73. AI-based Clinical Decision Support for Primary Care: A ... - OpenAI, dernier accès : juillet 30, 2025, https://cdn.openai.com/pdf/a794887b-5a77-4207-bb62-e52c900463f1/penda_paper.pdf
74. Ethical Issues for Autonomous Trading Agents - Strategic Reasoning Group, dernier accès : juillet 30, 2025, http://strategicreasoning.org/wp-content/uploads/2017/01/ethical-issues-autonomous.pdf
75. Ethics for AI and Robotics - Electrical Engineering and Computer Science, dernier accès : juillet 30, 2025, https://web.eecs.umich.edu/~kuipers/research/ethics/papers.html
76. Looking back, looking ahead: Humans, ethics, and AI, dernier accès : juillet 30, 2025, https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/18598/21504
77. Ethical content in artificial intelligence systems: A demand explained in three critical points, dernier accès : juillet 30, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10097940/
78. Multi-Value Alignment in Normative Multi-Agent System: An Evolutionary Optimisation Approach - MODeM 2023, dernier accès : juillet 30, 2025, https://modem2023.vub.ac.be/papers/MODeM2023_paper_18.pdf

79. Application-Driven Value Alignment in Agentic AI Systems: Survey and Perspectives - arXiv, dernier accès : juillet 30, 2025, https://arxiv.org/html/2506.09656v1
80. [2305.07366] Multi-Value Alignment in Normative Multi-Agent System: Evolutionary Optimisation Approach - arXiv, dernier accès : juillet 30, 2025, https://arxiv.org/abs/2305.07366
81. Value-Alignment Equilibrium in Multiagent Systems - CSIC Digital, dernier accès : juillet 30, 2025, https://digital.csic.es/bitstream/10261/235825/1/Value-Alignment%20Equilibrium%20in%20Multiagent%20Systems.pdf
82. The Labor Market At A Turning Point: AI Trends Through The Lens ..., dernier accès : juillet 30, 2025, https://dataconomy.com/2025/07/29/the-labor-market-at-a-turning-point-ai-trends-through-the-lens-of-katerina-andreeva/
83. Bridging the Experience Gap: Reducing Entry Barriers for New Graduates in the AI-Driven Labor Market - Goover, dernier accès : juillet 30, 2025, https://seo.goover.ai/report/202507/go-public-report-en-b2027993-7192-4636-bd65-5f772a1dbe57-0-0.html
84. Neurosymbolic AI as an antithesis to scaling laws | PNAS Nexus ..., dernier accès : juillet 30, 2025, https://academic.oup.com/pnasnexus/article/doi/10.1093/pnasnexus/pgaf117/8134151
85. Neurosymbolic AI. How hybrid models combining logic-based... | by Zaina Haider | Medium, dernier accès : juillet 30, 2025, https://medium.com/@thekzgroupllc/neurosymbolic-ai-2850dc2c7d8f
86. Unlocking the Potential of Generative AI through Neuro-Symbolic Architectures – Benefits and Limitations - arXiv, dernier accès : juillet 30, 2025, https://arxiv.org/html/2502.11269v1
87. [D] Why isn't more research being done in neuro-sumbolic AI direction? - Reddit, dernier accès : juillet 30, 2025, https://www.reddit.com/r/MachineLearning/comments/1ajrtug/d_why_isnt_more_research_being_done_in/
88. The Ethical Challenges of AI Agents | Tepperspectives, dernier accès : juillet 30, 2025, https://tepperspectives.cmu.edu/all-articles/the-ethical-challenges-of-ai-agents/
89. Trust is the new currency in the AI agent economy | World Economic ..., dernier accès : juillet 30, 2025, https://www.weforum.org/stories/2025/07/ai-agent-economy-trust/
90. Large Language Models as Theory of Mind Aware Generative ..., dernier accès : juillet 30, 2025, https://arxiv.org/pdf/2501.15355
91. arxiv.org, dernier accès : juillet 30, 2025, https://arxiv.org/html/2505.07087v2
92. A hybrid agent architecture integrating desire, intention and reinforcement learning | Request PDF - ResearchGate, dernier accès : juillet 30, 2025, https://www.researchgate.net/publication/220215355_A_hybrid_agent_architecture_integrating_desire_intention_and_reinforcement_learning

93. Agentic AI and Multiagentic: Are We Reinventing the Wheel? - arXiv, dernier accès : juillet 30, 2025, https://arxiv.org/pdf/2506.01463