Proposition de recherche doctorale (DIC-9411) : Architecture, formalisation et implantation du système sysCRED

Une approche hybride neuro-symbolique pour la crédibilité et le raisonnement en informatique cognitive

Dominique Loyer

Université du Québec à Montréal (UQAM)
Doctorat en informatique cognitive

Plan de la présentation

Introduction : La 3ème vague de l'IA

- **Oscillation historique** : Symbolisme (règles) $\leftrightarrow$ Connexionnisme (réseaux de neurones).
- **Convergence actuelle** : Nécessité de systèmes hybrides (neuro-symbolique).
- Allier la *généralisation neuronale* à la *rigueur symbolique* (Hitzler et al., 2025).

Le « Léviathan algorithmique »

- **Contexte** : Gouvernance automatisée par des algorithmes opaques (Hakim et al., 2025).
- **Bureaucratie numérique** : Vecteurs latents inintelligibles vs bureaucratie traditionnelle (règles écrites).
- **Déficit de crédibilité** des LLM (Large Language Models) :
  - Moteurs de corrélation statistique, pas de modèles causaux.
  - Hallucinations factuelles et grande assurance trompeuse.

Véracité vs crédibilité

Distinction épistémologique

1. **Véracité (Truthfulness)** : Correspondance énoncé/fait observable.
2. **Crédibilité (Credibility)** : Méta-propriété (fiabilité source, processus, cohérence) (Pan et al., 2025).

Problème

Les LLM compressent les sources et perdent le contexte. Les systèmes symboliques purs (GOFAI) sont fragiles face au web.

Hypothèse de recherche

Seule une **architecture hybride neuro-symbolique**, intégrant une ontologie de la crédibilité (Système 2) sur un modèle de langage perceptif (Système 1), permet d'atteindre la fiabilité requise.

Objectifs spécifiques (d'ici avril 2026)

- **Théorique (modélisation)** :
  - Formaliser une *Credibility Ontology* (biais, conflit d'intérêt, preuve, expertise).
  - Dépasser le binaire Vrai/Faux.
- **Technique (implémentation)** :
  - Concevoir **sysCRED** (System for Credibility and Reasoning utilizing Expert Dynamics).
  - Extraction neuro-symbolique et peuplement dynamique de graphe (GraphRAG).
- **Méthodologique (validation)** :
  - Double métrique : Précision (ML) + Qualité d'explication (cognitif).

Système 1 et système 2

Basé sur la *Dual Process Theory* (Kahneman) adaptée à l'IA (Yang et al., 2025).

Système 1 (intuitif)

- Rapide, parallèle, associatif.
- Réseaux de neurones profonds.
- Perception (Vision, NLP).
- → *Neural Interpreter*

Système 2 (délibératif)

- Lent, séquentiel, logique.
- Symboles explicites, règles.
- Planification, audit.
- → *Symbolic Auditor*

Ancrage des symboles (Symbol Grounding)

- **Défi** : Comment lier le symbole abstrait « Fake News » au texte réel ?
- **Approche sysCRED** : Vecteurs d'embedding des LLM comme pont vers l'ontologie.
- **Risque** : « Raccourcis de raisonnement » (Reasoning Shortcuts) (Marconato et al., 2025).
  - *Exemple* : Associer « Crédible » au style académique superficiel.
  - *Solution* : Régularisation logique stricte.

Positionnement : La 3ème vague (NeSy)

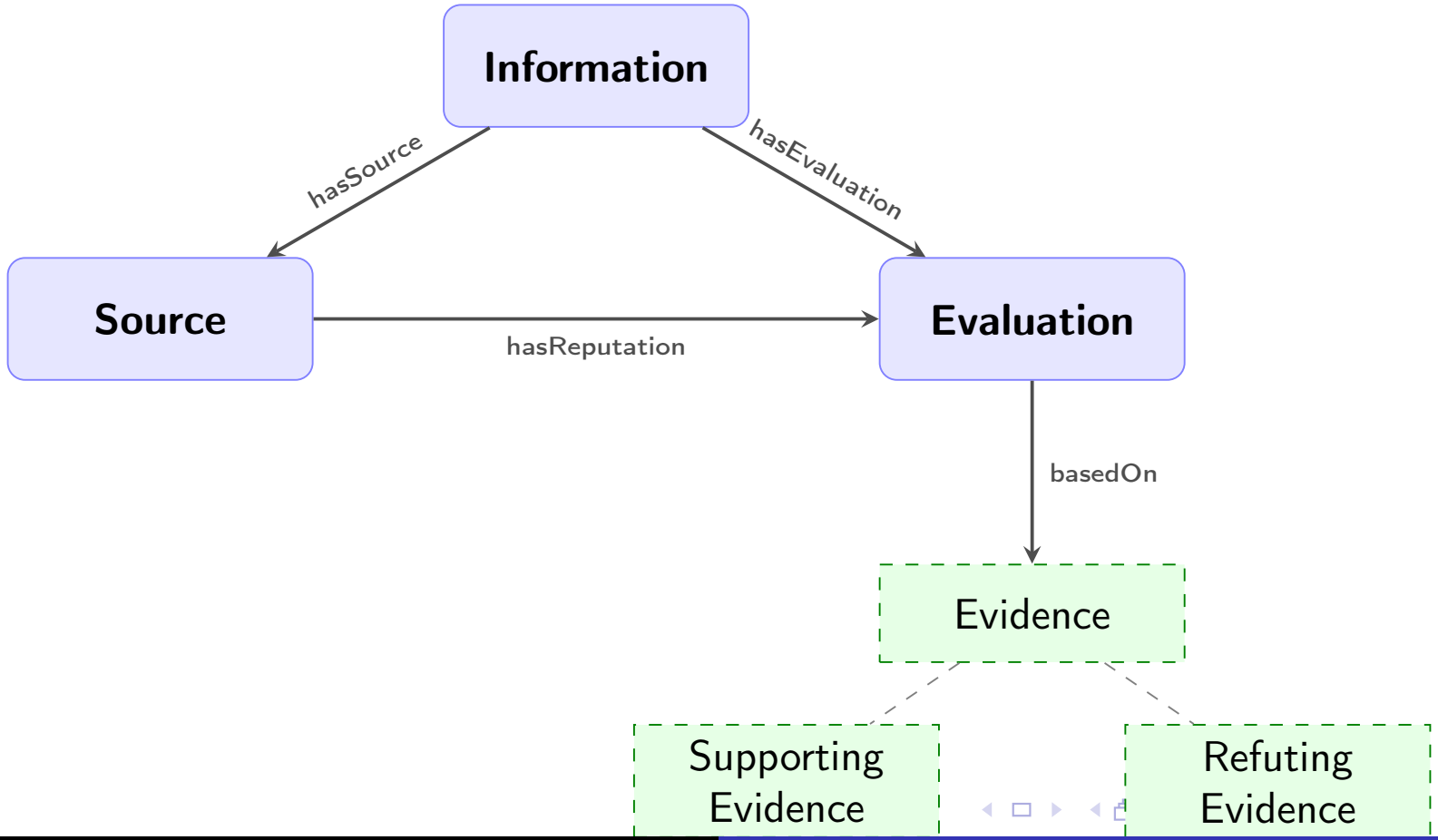
Trois grandes familles d'architectures (d'Avila Garcez and Lamb, 2023) :

1. **Hybride lâche (pipeline)** : Modules séquentiels.
2. **Intégration différentiable** : Logique dans la loss function (ex : Logic Tensor Networks).
3. **Programmation NeSy** : Induction de programmes neuronaux.

Choix sysCRED

Famille **Pipeline** : Seule approche garantissant une **auditabilité totale** (Zero Trust) nécessaire pour la lutte contre la désinformation.

Ontologie de crédibilité (classes principales)



IA de confiance (Trustworthy AI)

SysCRED s'inscrit dans le paradigme de l'**AI Assurance** (Sarker et al., 2023) :

- **Vérifiabilité** : Chaque décision possède une preuve dans le graphe.
- **Robustesse** : Résistance aux attaques adverses via le filtrage symbolique.
- **Interprétabilité** : Explications causales vs corrélations statistiques.

GraphRAG et Zero Trust

- **GraphRAG** (Retrieval-Augmented Generation sur Graphes) :
  - Récupère des sous-graphes, pas juste du texte.
  - Permet le raisonnement multi-sauts (Zhang et al., 2025).
- **Zero Trust AI** :
  - « Never Trust, Always Verify ».
  - Le module symbolique audite systématiquement le neuronal.

Design Science Research (DSR)

Création d'un artefact (sysCRED) pour résoudre un problème et générer des connaissances (Hevner et al., 2004; Peffers et al., 2007).

1. **Cycle de pertinence** : Besoins en explicabilité et traçabilité.
2. **Cycle de conception** :
   - Itération 1 : Pipeline Python/Turtle (Terminé).
   - Itération 2 : GraphRAG + Zero Trust (En cours).
   - Itération 3 : Optimisation et IHM.
3. **Cycle de rigueur** : Ancrage dans les standards (W3C, OWL) et théorie.

Vue d'ensemble : Le « Sandwich cognitif »

Architecture micro-services conteneurisée.

1. **Perception (S1)** : LLM fine-tunés (NER, Extraction Relations). Émet des assertions probabilistes.
2. **Le Pont (Bridge)** : Traduction Vecteur $\leftrightarrow$ Symbole (Grounding).
3. **Connaissances (graphe)** : Neo4j + RDFLib. Mémoire à long terme.
4. **Raisonnement (S2)** : Moteurs logiques (HermiT, Pellet). Règles SWRL.
   - *Règle Exemple* : Source satirique $\rightarrow$ Information fausse.

Flux de traitement (Workflow)

1. **Ingestion** : Texte/URL.
2. **Extraction neuronale** : Proposition de sous-graphe temporaire.
3. **Ancrage et GraphRAG** : Contextualisation via le Knowledge Graph global.
4. **Audit symbolique** : Vérification de cohérence logique (détection de contradictions).
5. **Synthèse** : Score de crédibilité + Explication causale en langage naturel.

Logique de pondération hybride (score)

Le score final est une aggrégation pondérée (Système 2 audite Système 1) :

- **Réputation de la source (symbolique)** : 25%
- **Google Fact-Check (Preuve externe)** : 20%
- **Cohérence et sentiment (neuronal)** : 30%
- **Entités et âge domaine** : 25%

Règle d'Or (Zero Trust)

Si la source est dans une **Liste Noire** connue → Score forcé à **0.0** (Veto), indépendamment de la qualité du texte.

Résultats du benchmark sysCRED v2.1

Résultats clés

- **Précision** : **NaN (0 Faux Positifs)**. Le système n'a validé aucune fausse information.
- **Rappel** : 0.00%. Le système est (trop) conservateur.
- **Interprétation** : Approche **Zero Trust** validée.

« La confiance ne se présume pas, elle se gagne par des preuves. »

Feuille de route (2026)

- **Phase 1 : Consolidation (Fév - Mars)**
  - Finalisation ontologie (Rhétorique, Biais).
  - Pipeline GraphRAG (LLM $\leftrightarrow$ Neo4j).
- **Phase 2 : Évaluation (Mars - Avril)**
  - Tests sur dataset LIAR.
  - Étude d'ablation (avec/sans moteur de règles).
  - Robustesse (Adversarial attacks).
- **Phase 3 : Finalisation (Avril)**
  - **Rédaction du rapport (3 avril 2026)**.
  - Publication (ISWC, AAAI).

Axes de recherche futurs (2026-2027)

1. **Axe 1 : Ancrage sémantique (Symbol Grounding)**
   - Utilisation des *Vector-Symbolic Architectures* (VSA) pour lier mathématiquement vecteurs et symboles.
2. **Axe 2 : GraphRAG interactif**
   - Interface conversationnelle pour « dialoguer » avec le graphe de preuves.
3. **Axe 3 : Ontologie dynamique**
   - Modélisation des campagnes de bots et de la désinformation temporelle.

Conclusion

- Réponse au *Léviathan Algorithmique* par une approche hybride rigoureuse.
- **sysCRED** : L'intuition probabiliste soumise à la vérification logique.
- Validité théorique (DSR) et pertinence sociétale (Désinformation).
- Vers une IA qui rend compte de ses raisonnements.

Références I

d'Avila Garcez, A. S. and Lamb, L. C. (2023). Neurosymbolic ai : The 3rd wave. *Artificial Intelligence Review*, 56.

Hakim, S. B., Adil, M., Velasquez, A., Xu, S., and Song, H. H. (2025). Neuro-symbolic ai for cybersecurity : A survey. *arXiv preprint arXiv :2509.06921*.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1) :75–105.

Hitzler, P. et al. (2025). Neuro-symbolic ai survey 2024-2025. *arXiv preprint arXiv :2501.05435*. v2 revised Apr 2025.

Références II

Marconato, E. et al. (2025). Symbol grounding in neuro-symbolic ai : A gentle introduction to reasoning shortcuts. *arXiv preprint arXiv :2510.14538*.

Pan, J. Z. et al. (2025). Large language models and knowledge graphs : Opportunities and challenges. *arXiv preprint arXiv :2504.07640*.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3) :45–77.

Références III

Sarker, M. K. et al. (2023). Neuro-symbolic methods for trustworthy ai : a systematic review. *Journal of Neuro-Symbolic AI*.

Yang, X.-W. et al. (2025). Neuro-symbolic artificial intelligence : Towards improving the reasoning abilities of large language models. *arXiv preprint arXiv :2508.13678*.

Zhang, Y. et al. (2025). A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv :2501.13958*.