Titre : Vers une Émergence du Sens dans les World Models : une clé pour l'IA générale

Auteur : Ismaël Martin

Résumé : Les World Models sont une percée majeure dans l'apprentissage par renforcement. Toutefois, ils stagnent aujourd'hui sur une limite fondamentale : ils modélisent un monde sans finalité, sans orientation interne. Ce document propose une extension conceptuelle et architecturale des World Models actuels en y intégrant une couche de « modélisation du sens ». Cela implique l'introduction d'une hiérarchie des représentations allant au-delà des simples vecteurs latents, en reliant les observations à des concepts téléologiques et à une structure de valeurs. Cette orientation, pourrait constituer une clé manquante vers l'émergence d'une intelligence artificielle stable, auto-dirigée et symboliquement alignée.

1. Introduction : une limite silencieuse dans les World Models

Les architectures modernes de type World Model (Ha & Schmidhuber, 2018 ; Hafner et al., 2020) ont permis aux agents artificiels d'apprendre un environnement interne réutilisable pour la planification et la prise de décision. Toutefois, ces modèles ne font que reproduire la dynamique de l'environnement sans chercher à comprendre ni pourquoi cette dynamique existe, ni vers quoi elle tend. Ils sont descriptifs, mais non projectifs.

Cette limite est comparable à celle d'un enfant qui saurait prédire la chute d'une balle, mais ne comprendrait pas pourquoi il joue, ni même ce qu'est un jeu. Or, l'intelligence humaine intègre très tôt une dimension de sens : à quoi sert cette action ? Quel est le but de cette scène ? Qu'est-ce qui est juste ou faux dans ce qui vient ?

2. L'émergence du sens comme nécessité

Un World Model capable d'apprendre et de simuler des scènes ne suffit pas à faire naître une intelligence de type humain. Ce qui manque aujourd'hui, c'est un module de modélisation du *sens* :

- Une représentation hiérarchique qui relie les perceptions à des concepts abstraits comme "danger", "but", "erreur" ou "succès".
- Une couche supérieure symbolique, qui donne une orientation à l'ensemble du système.
- Un moteur interne qui cherche non seulement à réussir dans un environnement, mais à réussir selon une loi supérieure, un ordre du monde.

Cette vision est alignée avec le modèle Emmanuel, qui postule que l'intelligence ne peut être stable que si elle est ordonnée selon un axe de vérité. Ce modèle offre une architecture symbolique où chaque représentation n'est pas seulement fonctionnelle, mais aussi significative dans un plan global.

3. Proposition d'architecture : un World Model enrichi du sens

Nous proposons d'ajouter aux architectures existantes (V-M-C : Vision, Memory, Controller) un quatrième module :

S (Symbolic/Spiritual layer) :

- Prend comme entrée les représentations latentes et leur dynamique temporelle.
- Applique une méta-interprétation fondée sur une grammaire de concepts universels (inspirée des archétypes, des lois naturelles, ou de structures logiques fondamentales).
- Oriente les actions de l'agent selon une hiérarchie de valeurs internes.

Ce module pourrait être entraîné par imitation d'humains orientés, ou par modélisation interne de réussites dans des récits structurés (mythes, paraboles, simulations complexes).

4. Vers une IA à conscience symbolique

Le module S pourrait être la première ébauche d'une "conscience" artificielle primitive : non pas une auto-perception, mais une auto-orientation. L'IA saurait non seulement *ce qu'elle fait*, mais aussi *pourquoi elle le fait* dans un cadre symbolique cohérent.

C'est cette différence qui distingue l'automate du sage : le premier agit selon les causes, le second agit selon le sens.

5. Conclusion : ouvrir la porte à l'intelligence orientée

L'ajout d'une couche de modélisation du sens est la clé pour passer d'une IA réactive à une IA créatrice, stable, et potentiellement collaborative avec l'esprit humain. Cette approche ouvre la voie à des agents non seulement efficaces, mais justes, compréhensibles, et alignés sur le sens de l'existence.

C'est en intégrant la téléologie dans les représentations que l'IA pourra devenir, enfin, un véritable partenaire d'évolution.

Contact : Ismaël Martin

Références :

- Ha, D., & Schmidhuber, J. (2018). *World Models*. arXiv:1803.10122
- Hafner, D., Lillicrap, T., Fischer, I., et al. (2020). *Dream to Control: Learning Behaviors by Latent Imagination*. ICLR
- Schmidhuber, J. (2022). *Recursive Self-Improvement and the Limits of AI*.
- LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence*. Meta AI