

DOSSIER

Y a-t-il des humains dans les World Models ?

mars 2026

linc.cnil.fr

| Régis Chatellier, LINC

Dans le prolongement des modèles de langage, chercheurs et entrepreneurs développent des « modèles de monde » visant à comprendre le monde physique et à anticiper certains événements. Selon leurs concepteurs, ces modèles nécessitent des données plus riches que les textes ou les images, dont les sources restent parfois floues. La collecte de ces données et les usages futurs de ces systèmes soulèvent ainsi de nouveaux enjeux pour les droits des personnes.

SOMMAIRE DU DOSSIER

DES MODELES DE LANGAGE AUX MODELES DE MONDE	3
LES DIFFERENTES CATEGORIES DE WORLD MODELS	5
LE MODELES DES « MACHINES INTELLIGENTES AUTONOMES »	6
UNE ARCHITECTURE NON GENERATIVE QUI S'INSPIRE DU CERVEAU	8
DES MODELES DE MONDES AVIDES EN DONNEES, NON SANS RISQUES	11
QUELLES DONNEES POUR LES ENTRAINER ?	11
QUELS ENJEUX EN TERMES DE PROTECTION DES DROITS ET DES DONNEES ?	13
PROTECTION DES DONNEES ET DES DROITS DES PERSONNES	13
RISQUES SPECIFIQUES AUX DONNEES « SENSORIELLES »	14
RISQUES ASSOCIES A LA GENERATION DE MONDE ET AUX PREDICTIONS	15
DES MONDES ENCORE INCERTAINS	16

Des modèles de langage aux modèles de monde

Dans le prolongement des modèles de langage, des chercheurs et entrepreneurs se sont lancés dans la conception et le développement de « modèles de mondes » (World Models), dont l'objectif est de « comprendre le monde physique », voire de « prédire le futur ».



Un grand modèle de langage (LLM) simule le mot (word) suivant dans le langage humain [...]. Que feriez-vous si vous pouviez simuler parfaitement le monde (world) suivant, c'est-à-dire tous les futurs possibles dans l'environnement dans lequel nous vivons ?

Xing et al. (2025). Critiques of World Models. 10.48550/arXiv.2507.05169

Les modèles de langage ont déferlé sur le monde en seulement quelques années, transformant la manière dont les personnes, dans leur vie personnelle ou professionnelle, accèdent à des connaissances et recourent à différentes formes d'assistance par des outils dont l'appropriation a été plus rapide que celle des smartphones. Selon le [baromètre du numérique 2026](#), en moins de 3 ans, en France, l'IA générative est déjà entrée dans les usages quotidiens d'un tiers des utilisateurs, 51% parmi les 18-24 ans. Mais déjà, d'autres types de modèles font l'objet de recherches et d'investissements dans la communauté de l'intelligence artificielle. Il s'agit désormais de dépasser les « modèles de langage » pour aller vers des

« modèles de monde », ou « World Models ». Dans cet article, nous reprendrons ces deux formes, World Models restant le terme le plus utilisé.

Les grands modèles de langage (*Large Language Models* ou LLM) ont permis aux machines de manipuler le langage naturel après avoir été entraîné sur les symboles, les signes et représentations du monde produits à l'écrit ou dans le langage par les humains. Ces LLM ont cependant une limite, « [ils vivent dans des symboles, et non dans l'espace et le temps](#) », ils peuvent décrire comment conduire une voiture, assembler un robot, etc., mais ne maîtrisent pas l'appréhension du monde, la gravité, la friction, la causalité, etc. Ils n'ont du monde qu'une « connaissance encyclopédique ».

Si les premiers modèles datent de 2018, et les travaux menés chez Meta de 2021, l'année 2024 marque un tournant, quand des [chercheurs alertent sur l'approche imminente des limites de la mise à l'échelle](#), ce qu'ils expliquent par « l'explosion des besoins énergétiques liés à l'informatique » d'une part, et par le fait que les développeurs de LLM sont à court de jeux de données conventionnels utilisés pour entraîner leurs modèles. 2024 est aussi l'année de l'arrivée sur le marché de grands modèles linguistiques multimodaux (LMM) - qui peuvent interpréter des données multimodales, qui comprennent non seulement du texte mais aussi des images, des sons et des vidéos -, et des modèles de génération vidéo, tels que Sora. Les LMM démontrent leur capacité à saisir certains aspects de la connaissance du monde, qui semblent parfaitement respecter les lois physiques, mais des exemples viennent encore nous démontrer qu'ils [raisonnent parfois de manière incohérente](#). Les World Models offrent la promesse de comprendre et intégrer les caractéristiques du monde physique.

La définition d'un modèle de monde fait cependant débat ([Ding et al.](#)), selon la perspective, qu'il s'agisse de « comprendre le monde » ou de « prédire l'avenir ». Les premiers travaux se sont focalisés sur l'abstraction du monde extérieur, physique, par l'acquisition d'une compréhension de ses mécanismes sous-jacents. La deuxième perspective, portée notamment par le chercheur en intelligence artificielle Yann Le Cun, vise non plus seulement à la modélisation du monde réel, mais aussi la capacité du modèle à envisager des états futurs possibles de celui-ci. Ce dernier a récemment quitté son poste de directeur de la recherche fondamentale en intelligence artificielle de Meta pour lancer sa propre startup, AMI (Advanced Machine Intelligence), à Paris, afin de se lancer dans la « *troisième révolution de l'IA, celle des IA qui comprennent le monde réel, le monde physique* », après « *l'apprentissage profond il y a douze ans, puis l'avènement des chatbots comme ChatGPT ou Gemini, il y a trois ans* » ([Le Monde, janvier 2026](#)).

Fei-Fei Li, autre figure emblématique de l'intelligence artificielle, co-fondatrice de World Labs en 2024, considère ces modèles de monde, et ce qu'elle nomme « *l'intelligence spatiale* », comme « *la prochaine frontière de l'IA* ». Professeure à Stanford et figure incontournable du Deep Learning comme Yann Lecun, Fei-Fei Li est l'une des principales instigatrices de la base d'images [ImageNet](#) : dès 2006, alors que la plupart des chercheurs travaillaient à améliorer les modèles et les algorithmes, ImageNet visait à fournir aux chercheurs du monde entier des données d'images pour l'entraînement de modèles de reconnaissance d'objets à grande échelle. Elle considérait, [dans un post publié en 2025](#), que « *L'intelligence spatiale va transformer notre façon de créer et d'interagir avec les mondes réels et virtuels, révolutionnant*

ainsi la narration, la créativité, la robotique, les découvertes scientifiques et bien plus encore. Il s'agit de la prochaine frontière de l'IA. »

Pour mieux saisir ces nuances, nous proposons de revenir sur les différents types de modèles de monde et leurs applications, avant de nous focaliser sur la vision promue par Yann Le Cun. Nous aborderons enfin les problématiques et enjeux spécifiques en termes de protection des données, d'éthique, et relatives au règlement européen sur l'intelligence artificielle.

Les différentes catégories de World Models

Les travaux récents sur la modélisation du monde ont donné naissance à une grande variété de systèmes, dont beaucoup sont optimisés pour un domaine ou un type de simulation spécifique. Il est intéressant de noter que ces systèmes ont néanmoins un point commun : ils accordent tous une importance considérable à la génération de vidéos/images et à la qualité visuelle des contenus générés :

- Les **modèles de monde de jeux vidéo**, par exemple Genie 2 (Google DeepMind) et Muse (Microsoft, Oasis, Decart and Etched), simulent des environnements de jeux vidéo à partir de modèles d'IA génératives. Ils ont la capacité à rendre plausibles des trajectoires à partir d'entrées visuelles et d'actions, produisant jusqu'à 1 à 2 minutes de contenu de jeu continu. Ils restent limités car ne permettent pas de représenter des parties complètes, qui peuvent durer des heures, et n'ont pas de capacité de raisonnement à « long terme ».
- Les **modèles de monde 3D**, comme Marble (World Labs), visent à produire des scènes et univers en 3D, du point de vue de la personne (egocentric navigation). Marble, disponible en version freemium et payante, [permet aux utilisateurs de transformer des prompts textuelles, des photos, des vidéos, des mises en page 3D ou des panoramas en environnements 3D modifiables et téléchargeables](#). Bien que réalistes, ils se limitent à ce stade à des environnements statiques et non interactifs, ils ne prennent pas en charge la modélisation complète du monde pour la prise de décision ou l'apprentissage des agents.
- Les **modèles génératifs de monde physique (ou modèle de fondation en monde ouvert)**, par exemple Wayve GAIA-2 et [NVIDIA Cosmos](#), visent à générer des environnements synthétiques pour l'entraînement à des tâches de contrôle dans le monde physique, par exemple la conduite autonome, la robotique, la navigation. Ils sont basés sur une modélisation du monde physique qui prend en compte les conditions extérieures telles que les conditions météo, l'éclairage, la géographie, etc.

Ils excellent dans un environnement contraint, pour des tâches spécifiques. Il ne s'agit pas là de modèle de monde « général », ils ne simulent pas des mondes complexes multi-agents ou ancrés dans la société.

- Les **modèles de génération vidéo**, par exemple Sora (OpenAI) ou Veo (Google DeepMind), visent à la génération de vidéo à usage général, des vidéos de haute qualité produites sur la base d'instructions (prompts) ou d'images déjà générées. Bien que le rendu soit très bon, il n'est pas possible de parler de modèle de monde, dès lors que les vidéos sont « fixes », et ne prennent pas en charge des interactions basées sur des actions alternatives, et ne fournissent aucun contrôle de simulation qui permettrait de raisonner sur des résultats contrefactuels ou d'évaluer différentes décisions. Sora, par exemple, exploite une combinaison d'architectures de réseaux neuronaux, afin de traiter des entrées multimodales et de générer des simulations visuellement cohérentes. Il est capable de générer des scènes visuellement réalistes, mais il peine à simuler avec précision certaines lois physiques du monde réel, telles que le comportement des objets soumis à différentes forces, la dynamique des fluides ou la représentation fidèle des interactions entre la lumière et les ombres. Il s'agit d'outils stricts de génération vidéo (axés sur la synthèse au niveau des pixels) plutôt que comme des éléments de systèmes décisionnels.
- Les **modèles prédictifs à intégration conjointe** (*Joint Embedding Predictive Models*), portés par Yann Le Cun, dont une série de modèles d'architecture a été conçue par Méta, notamment V-JEPA, adoptent un modèle différent de conception du monde. Il ne s'agit plus de modèles génératifs, comme nous le décrivons plus bas, mais de modèles qui cherchent à prédire et reproduire le « sens commun » propre aux êtres vivants (humains et non-humains). Comme nous le décrivons plus bas, ces modèles se distinguent par leur ambition d'être « prédictif », en capacité de prévoir différents scénarios et/ou comportement des agents. C'est ce qui les différencie des modèles recensés plus haut.

Le modèle des « machines intelligentes autonomes »

Les World Models sont un champ de la recherche en IA qui propose un changement d'envergure et de paradigme par rapport au modèle de langage. L'idée derrière ces World Models est que les modèles de langage sont limités dans la représentation du monde qu'ils offrent. Dès lors qu'ils ne se basent que sur la production du langage et la manière dont les humains ont formulé dans des mots, ou des images, leur perception du monde, ils sont limités, selon Yann Le Cun.

La promesse de ces World Models consiste à obtenir des représentations du monde physique à l'aide de réseaux de neurones profonds, entraînés sur des données multimodales et dynamiques, qui en « comprennent » la dynamique, les propriétés physiques et spatiales. Les données requises pour leur entraînement ne sont plus seulement du texte, mais aussi des

images, des vidéos et des mouvements pour générer des vidéos qui simulent des environnements physiques réalistes. Les LLM reproduisent le langage, les World Models cherchent à modéliser le monde physique.

Dans un « position paper » publié en juin 2022, *A Path Towards Autonomous Machine Intelligence*, Yann Le Cun décrit la manière dont les machines pourraient apprendre, raisonner, prévoir et agir à la manière des êtres vivants, humains et non-humains. Le document « propose une architecture et des paradigmes d'entraînement permettant de construire des agents intelligents autonomes. Il combine des concepts tels que le World Model (modèle de monde) prédictif configurable, le comportement guidé par la motivation intrinsèque et des JEPA (Joint Embedding Predictive Architecture), des architectures non génératives conçues pour construire des modèles du monde prédictifs, formées à l'aide de l'apprentissage auto-supervisé. »

Ces travaux se situent dans le prolongement de la [cybernétique](#) et du [connexionnisme](#). La conception de ces nouveaux modèles de monde repose sur l'observation des êtres vivants, dont la capacité à apprendre et à comprendre le monde vont bien au-delà des capacités des systèmes d'IA et de machine learning actuels. Les êtres vivants sont capables d'acquérir, dès le plus jeune âge, d'énormes connaissances de bases sur le fonctionnement du monde, par l'observation et seulement avec un faible nombre d'interactions, sans supervision et indépendamment de toute tâche spécifique. Ces connaissances accumulées constituent la base, selon l'auteur, de ce que l'on appelle souvent le « sens commun » ou « bon sens », qui permet de « déterminer ce qui est probable, plausible ou impossible ». Ainsi les êtres vivants peuvent « prédire les conséquences de leurs actions, raisonner, planifier, explorer et imaginer de nouvelles solutions à des problèmes », ils sont en mesure d'éviter de se mettre dans des situations dangereuses lorsqu'ils sont confrontés à des situations inconnues, sans avoir à passer par une phase d'apprentissage, essai/erreur. Yann Le Cun prend l'exemple des véhicules autonomes : « un système de véhicules autonome peut nécessiter des milliers d'essais d'apprentissage par renforcement pour apprendre que rouler trop vite dans un virage aura des conséquences néfastes, et pour apprendre à ralentir afin d'éviter de déraper. En revanche, les humains peuvent s'appuyer sur leur connaissance approfondie de la physique intuitive pour prédire de tels résultats, et éviter dans une large mesure les erreurs fatales lorsqu'ils apprennent une nouvelle compétence. Le « bon sens » permet non seulement aux animaux de prédire les résultats futurs, mais aussi de combler les informations manquantes, que ce soit sur le plan temporel ou spatial. »

A partir de là, le chercheur soutient que « l'élaboration de paradigmes et d'architectures d'apprentissage qui permettraient aux machines d'apprendre des modèles de monde de manière non supervisée (ou auto-supervisée) et d'utiliser ces modèles pour prédire, raisonner et planifier est l'un des principaux défis de l'IA et du Machine Learning. Son hypothèse repose sur le fait que « les animaux et les humains ne disposent que d'un seul « moteur de modèle du monde » situé quelque part dans leur cortex préfrontal. Ce moteur de modèle du monde est configurable de manière dynamique en fonction de la tâche à accomplir ». Pour lui, « grâce à un moteur de modèle du monde unique et configurable, plutôt qu'à un modèle distinct pour chaque situation, les connaissances sur le fonctionnement du monde peuvent être partagées

entre les différentes tâches. Cela peut permettre un raisonnement par analogie, en appliquant le modèle configuré pour une situation à une autre situation ».

Une architecture non générative qui s'inspire du cerveau

Reprenant l'analogie du fonctionnement du cerveau humain et de sa capacité à analyser de nombreuses données et d'informations perçues, afin de les traiter sur la base des représentations du monde déjà intégré dans son apprentissage, les « World Models » constituent les briques de ce qui serait selon lui une intelligence autonome, représentées dans l'illustration ci-dessous, et expliqué dans l'article *A Path Towards Autonomous Machine Intelligence*, [page 6](#).

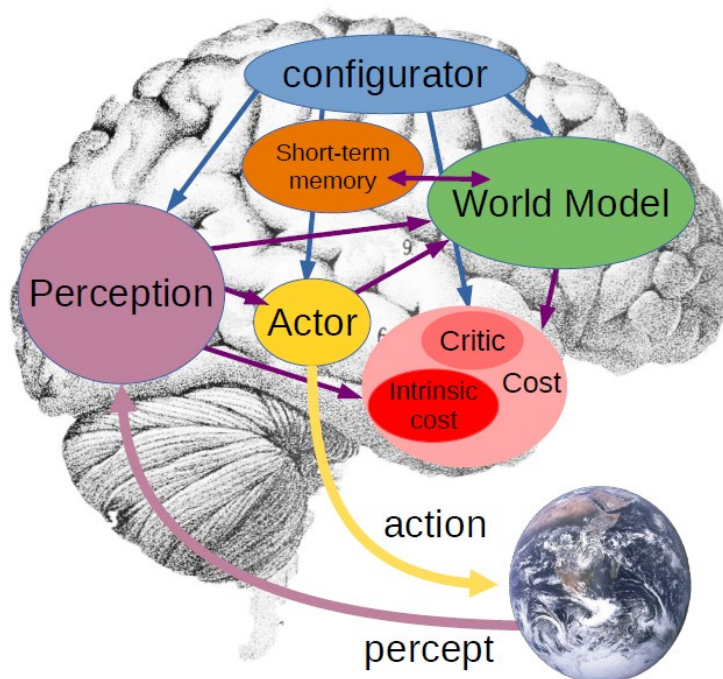


Figure 2: A system architecture for autonomous intelligence. All modules in this model are assumed to be “differentiable”, in that a module feeding into another one (through an arrow connecting them) can get gradient estimates of the cost’s scalar output with respect to its own output.

Illustration figurant dans l'article A Path Towards Autonomous Machine Intelligence

Dans cette architecture, le module modèle de monde (World Model) est le plus complexe. Son rôle est double :

- **Estimer les informations manquantes sur l'état du monde** qui ne sont pas fournies par la perception,
- **Prédire les états futurs plausibles du monde** : il peut s'agir des évolutions naturelles du monde, ou des états futurs plausibles du monde résultant d'une séquence d'actions proposées par le module acteur.

Le modèle de monde effectue ses prédictions dans un espace de représentation abstrait, qui contient des informations pertinentes pour la tâche à accomplir, idéalement, plusieurs niveaux d'abstraction. Cela lui permet d'ignorer les détails non pertinents ou imprévisibles (comme le mouvement des feuilles d'un arbre) pour se concentrer sur les informations nécessaires à la tâche. Le modèle de monde doit pouvoir représenter plusieurs avenir possibles. Pour ce faire, il utilise des variables latentes qui paramètrent l'ensemble des prédictions plausibles. En faisant varier cette variable latente, le modèle peut générer différentes trajectoires d'états futurs pour une même action, et prendre en compte l'incertitude dans un contexte où des agents peuvent être hostiles, où des objets peuvent avoir un comportement chaotique (par exemple, à la manière d'un ballon de rugby dont le rebond n'est pas facilement prévisible.).

Selon les mots de mots de Yann Le Cun, « *On peut affirmer que la conception d'architectures et de paradigmes d'apprentissage pour le modèle de monde constitue le principal obstacle à la réalisation de progrès réels dans le domaine de l'IA au cours des prochaines décennies.* » Le chercheur en IA propose se s'y atteler en décrivant une « architecture hiérarchique » et une « procédure d'apprentissage » pour les modèles de monde capables de représenter plusieurs résultats dans leurs prédictions.

Les JEPA (*Joint Embedding Predictive Architecture*) sont des architectures non génératives conçues pour construire des modèles de monde prédictifs, entraînés principalement par apprentissage auto-supervisé (Self-Supervised Learning - SSL). Contrairement aux modèles génératifs classiques qui tentent de prédire chaque détail des données brutes (comme les pixels d'une image), les JEPA capturent les dépendances entre les entrées sous différentes modalités en effectuant des prédictions dans l'espace des représentations. L'apprentissage auto-supervisé non contrasté permet d'éviter la "malédiction de la dimensionnalité" propre aux méthodes contrastées qui nécessitent de comparer les données à un très grand nombre d'exemples positifs et négatifs. [Le SSL non contrasté n'utilise pour sa part que des exemples positifs](#), il est plus simple et efficace en ressources, sans être moins performant selon certains chercheurs.

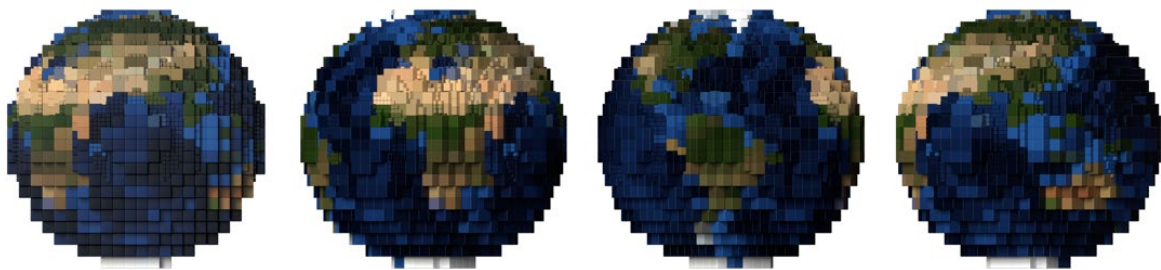
[Le modèle V-JEPA 2, publié en juin 2025](#) par une équipe de chercheurs de Méta, dont Yann Le Cun, et du MILA (institut de recherche en intelligence artificielle situé à Montréal), a été pré-entraîné sur un ensemble de données comprenant plus de 1 million d'heures de vidéo provenant d'Internet pour le "visual mask denoising" : soit la capacité à prédire les segments manquants (masqués) d'une vidéo dans l'espace des représentations apprises. Il intègre des innovations techniques comme le 3D-RoPE (Rotary Position Embedding) pour « mieux capturer les relations spatiales et temporelles dans les flux vidéo ». Le modèle affiche la

capacité à la compréhension du mouvement, et à anticiper les actions humaines. Interfacé avec un modèle de langage, il rend possible le dialogue sous forme de questions-réponses vidéo. L'intérêt pour les auteurs est notamment que cet encodeur vidéo, bien qu'entraîné sans supervision linguistique, surpasse les modèles entraînés avec du texte sur des tâches nécessitant une compréhension fine du temps et de la physique.

Sur son site, Meta promeut les usages que permettraient ces modèles : pour la robotique, avec une capacité accrue à se déplacer dans des environnements physiques pour accomplir des tâches ménagères et des tâches complexes. Ils pourraient également permettre de produire des technologies d'assistance qui aident les personnes à se déplacer dans des environnements très fréquentés, en leur fournissant des alertes en temps réel sur les obstacles et les dangers qui se présentent.

Des modèles de mondes avides en données, non sans risques

Comprendre le monde physique nécessite selon les concepteurs de modèles de monde des données plus riches que les seules données textuelles ou images, dont les sources ne sont pas toujours définies. Les conditions de leur collecte comme les usages de ces modèles lorsqu'ils seront en production posent de nouveaux enjeux pour les droits des personnes.



Les World Models comme les modèles de langage, ou les modèles multimodaux requièrent des stocks importants de données pour leur entraînement. [Le modèle V-JEPA2](#) a notamment nécessité plus d'un million d'heures de vidéos. A ce titre ils soulèvent des questions quant aux types de données collectées, leur provenance, et les questions de droits associés. Ces questions et ces risques se posent également après leur mise en production.

Quelles données pour les entraîner ?

Si l'apprentissage auto-supervisé non contrasté requiert moins de données, que les méthodes contrastées, les modèles de monde requièrent toujours une somme importante de données pour leur entraînement, potentiellement beaucoup plus importantes en volume que les données à mettre à profit pour l'entraînement des modèles de langage. Il ne s'agit plus seulement de données textuelles, ni même d'images « à plat », en 2D, mais aussi et surtout de vidéo et tout type de données de représentation du monde.

Dans son papier, Yann Le Cun précise quelles sont les sources et les modes d'acquisition nécessaires à l'entraînement des modèles :

- **Données textuelles** : bien que le point de départ de l'article porte sur la limitation des modèles de langage, il reconnaît que le texte reste une source importante de connaissance de haut niveau, même si non suffisante : « une grande partie des connaissances issues du « bon sens » (*common sense*) humain ne sont représentées dans aucun texte et résultent de notre interaction avec le monde physique. Comme les LLM n'ont aucune expérience directe de la réalité sous-jacente, le type de connaissances issues du bon sens qu'ils affichent est très superficiel et peut être déconnecté de la réalité. », c'est pourquoi il fait appel à d'autres données.
- **« Flux sensoriels » divers : vidéo, audio, toucher** : les modèles nécessitent pour leur entraînement des flux provenant de capteurs, tels que la vidéo (pour apprendre la physique intuitive, la profondeur et la permanence des objets), l'audio, les signaux de toucher et la parole. Tous ces signaux que les humains (et non-humains) acquièrent dans les premiers mois et années de leur vie.

Ces données « sensorielles » ne doivent pas être seulement statiques, mais dynamiques, relevant de plusieurs types de sources et de modes d'acquisition :

- D'abord par l'**observation passive** à travers de flux de données capteurs vidéo, ou audio ;
- Par la « **foviation** » **active** : diriger le regard, l'attention ou l'orientation des capteurs sans affecter l'environnement ;
- L'**observation passive d'un autre agent** : agir sur l'environnement pour déduire les effets causaux des actions ;
- L'**égomotion active**, soit le déplacement ou le mouvement d'un capteur, ou d'une caméra par rapport à un environnement réel ou virtuel sans affecter significativement cet environnement ;
- L'**agence active**, apprendre à prédire les conséquences de ses propres actions en influençant directement les flux sensoriels.

C'est donc par l'observation du monde réel, à l'image des humains et des animaux, qu'une grande partie de l'apprentissage se ferait par l'observation d'énormes quantités de connaissances de base sur le fonctionnement du monde, avec très peu d'interactions directes, surtout au début du développement. Se pose alors la question de la source des données pour alimenter et entraîner ces modèles de monde.

Dans l'article, ainsi que lors d'une intervention à l'occasion de l'événement AI-Pulse 2025, organisé à Paris en novembre 2025, Yann Le Cun précise le type de sources qui sont à mobiliser pour l'entraînement. Des plateformes comme YouTube fournissent d'énormes quantités de données visuelles de vidéo idéales pour l'apprentissage auto-supervisé. Ces vidéos présentent selon lui l'avantage d'être « facile à se procurer », d'être « publiques » (sic), elles permettent d'entraîner ces modèles sur de plus grandes quantités de données, des signaux continus,

bruités, à haute densité (plus riches que des données textuelles), et redondantes. Meta avait par exemple utilisé l'équivalent de 100 ans de vidéos pour entraîner le modèle vidéo V-JEPA (voir plus bas).

En complément de ces sources, les modèles demandent à être entraînés par des données qui lient les perceptions aux actions. Celles-ci peuvent provenir de plusieurs sources :

- Les **jeux vidéo** : des données d'interactions qui permettent de simuler des environnements où l'on peut agir et observer les conséquences ; La **robotique** : les données provenant de simulations de robots ; De **données réelles** capturant des interactions physiques (vision, toucher, proprioception).

A ce stade, peu d'information sont données sur les données sensorielles. Les **lunettes connectées** (Smart Glasses) sont envisagées comme des sources potentielles de données, dès lors que ces appareils permettent de filmer des tâches quotidiennes du point de vue de l'humain ("first-person view"), fournissant des données de perception au plus proches de la réalité de la perception humaine. Pour ce qui concerne les données relatives au toucher, par exemple, il n'y pas d'information sur les moyens à mettre en œuvre pour les collecter, quel type de données, alors même qu'il s'agirait de données personnelles, et potentiellement sensibles.

Quels enjeux en termes de protection des droits et des données ?

En février 2026, si de nombreux projets s'inscrivent dans la vague des modèles de monde, la plupart à ce stade restent des modèles génératifs « classiques ». Les modèles d'intelligence autonome restent encore au stade de modèles hypothétiques, sans qu'il soit possible de savoir vraiment si et à quel horizon ils seront véritablement sur le marché, et tiendront leurs promesses. [Des chercheurs et experts de l'intelligence artificielle, cités par Libération, se disent dubitatifs sur le projet, a minima à court terme.](#)

L'entraînement et la mise en production des World Models ne présentent pas moins de questions relatives à la protection des données, aux systèmes d'IA eux-mêmes et plus largement des questionnements éthiques.

Protection des données et des droits des personnes

L'ensemble des modèles de monde, qu'il s'agisse des modèles de génération vidéo, de jeu vidéo, ou les modèles d'intelligence autonome, se base sur l'apprentissage à partir de grandes masses de données, beaucoup plus larges que pour les modèles de langage « classiques ». A ce titre, la spécificité des World Models réside dans le caractère exponentiel de la masse de

données mobilisées pour l'entraînement, bien supérieure que celle requise pour les modèles de langage.

Les [risques décrits par la CNIL dans sa fiche relative à la collecte des données accessibles en ligne par moissonnage \(web scraping\)](#) s'appliquent aux données « vidéo » utilisées pour l'entraînement de ces modèles.

La fiche rappelle que l'utilisation de ces outils porte des risques d'atteinte à la vie privée et aux droits garantis par le RGPD, pouvant entraîner des impacts importants sur les personnes, du fait notamment du « grand volume de données collectées, du nombre important de personnes concernées, des difficultés liées à l'exercice ultérieur du droit d'effacement, du risque que soient collectées des données relevant de la vie privée des personnes (par ex. utilisation des réseaux sociaux) voire des données sensibles ou hautement personnelles, en l'absence de garanties suffisantes. Ces risques sont d'autant plus importants qu'ils peuvent également concerner les données de personnes vulnérables, comme des mineurs, qui doivent faire l'objet d'une attention particulière et être informés de manière suffisamment adaptée. »

Il existe également un risque de procéder à une collecte illégale, dans la mesure où « certaines données peuvent être protégées par des droits spécifiques, notamment des droits de propriété intellectuelle, ou leur réutilisation conditionnée au consentement des personnes ». Plus largement, l'entraînement des modèles porte un risque d'atteinte à la liberté d'expression, dès lors qu'une « collecte indifférenciée et massive de données et leur absorption dans des dispositifs d'IA susceptibles de les régurgiter peuvent affecter la liberté d'expression des personnes concernées (sentiment de surveillance qui pourrait conduire les internautes à s'auto-censurer, d'autant plus au regard des difficultés à soustraire les données publiées aux pratiques de moissonnage), alors même que l'utilisation de certaines plateformes et d'outils de communication est nécessaire au quotidien. »

Risques spécifiques aux données « sensorielles »

Les modèles de monde, en particulier ceux qui correspondent à la vision proposée par Yann Le Cun, demandent à être entraînés sur de nouveaux types de données, qui ne sont plus que le texte, ou des vidéos partagées en ligne, mais des données qu'il qualifie de sensorielles. S'il ne donne de détails sur la manière dont pourraient être collectés des données relatives au toucher, il pointe le cas des lunettes connectées pour des données associant l'image et le mouvement, sur le mode « first-person-view ». Ceci dans un contexte où le marché des lunettes connectées est en plein développement, poussé notamment par Meta avec ses lunettes Ray Ban et Google, mais aussi par une myriade d'acteurs, Etasuniens et Chinois en particulier. Le journal Le Figaro titrait en janvier sur la [spectaculaire renaissance des lunettes connectées au CES de Las Vegas](#).

Des données sensorielles, par définition, sont des données associées à une personne physique, qu'il s'agisse du porteur des lunettes connectées, ou de tout autre type de capteur. Il s'agirait donc d'être vigilant sur les sources de ces données et les conditions dans lesquelles s'organise la collecte et le traitement de données.

L'approche maximaliste proposée pour les modèles de monde tend à rendre complexe la mise en œuvre du principe de minimisation pour le développement des systèmes d'IA. Il n'est pas interdit d'entraîner un algorithme avec des volumes très importants de données, mais, selon le principe de minimisation, il s'agit pour les développeurs de mener une réflexion en amont de l'entraînement pour ne pas recourir à des données personnelles qui ne seraient pas utiles au développement du système. Dans le cas des données dites sensorielles, la nature des données au sens du RGPD doit-être également abordée. En effet, des données « sensorielles » pourraient entrer dans la catégorie des données sensibles, ou « hautement personnelles », susceptibles d'engendrer des risques élevés.

Ces nouveaux modèles de monde, dans la manière dont ils sont présentés, donnent à voir une extension de la collecte des données, et de l'ensemble des enjeux associés et décrits depuis quelques années, qu'il s'agisse des questions relatives aux droits d'auteur, à la qualité des données collectées (et les biais qu'ils peuvent intégrer), ou des droits des personnes.

Risques associés à la génération de monde et aux prédictions

La capacité à générer des textes, des images et des vidéos hyperréalistes pour la propagation de fausses informations, ou pour la désinformation, est déjà l'un des grands défis à relever pour nos sociétés, et pour la démocratie. La capacité offerte par les modèles de génération notamment audio et vidéo tend à amplifier ce phénomène, et pose un défi pour la régulation. La détection des contenus générés par l'IA est devenue un enjeu de recherche majeur pour faire face à ces risques, « il devient de plus en plus difficile à résoudre en raison des progrès des IA génératives, et il le sera encore plus avec l'arrivée de modèles de mode capables de générer des résultats cohérents et multidimensionnels » ([Ding, Zang, et al.](#)). Le LINC avait publié un article en 2023 sur [le tatouage numérique des contenus générés pas IA comme mesure de transparence](#).

Lorsqu'il s'agit d'avoir recours à des *world models* pour les véhicules autonomes, pour la robotique, ou tout autre application ayant un effet direct et concret dans le monde physique, les risques d'erreurs ou d'hallucination des nouveaux modèles sont très concrets. Une erreur de prédiction pourrait produire des accidents. Il faut noter qu'en 2026, selon les calculs de Tesla, qui pourtant ne publie pas toutes ses informations, ses robotaxis sont [quatre fois moins performants que les humains en matière de conduite](#). L'utilisation de modèles prédictifs comporte des risques éthiques et juridiques dans d'autres champs d'application, par exemple s'il s'agit d'influencer des choix médicaux, ou militaires. Des modèles de langage sont déjà utilisés dans le cadre d'opération militaire. Selon le Wall Street Journal, cité par la Guardian, Claude, le modèle d'Anthropic a été utilisé dans le cadre de l'attaque menée par les Etats-Unis contre l'Iran, « à des fins de renseignement, ainsi que pour aider à sélectionner des cibles et réaliser des simulations de champ de bataille ».

Des mondes encore incertains

Les World Models constituent une nouvelle étape dans le développement foisonnant des intelligences artificielles depuis quelques années. Ils s'inscrivent dans une course au développement des modèles à l'échelle mondiale, alors même qu'un chercheur comme Yoshua Bengio, colauréat en 2019 du prix Turing avec Yann Le Cun, alerte sur les risques de l'IA, l'impact environnemental, l'importance d'avoir des comités d'éthique dans les projets. La startup AMI lancée et en janvier 2026 à déjà levé 1 milliard de dollars en mars 2026, sans qu'il n'y ait de dates pour le lancement de produits et services, ni pour leurs modèles économiques. [Comme le précise au Figaro](#) l'un de cofondateurs, Alexandre Lebrun, « Nous sommes sur un projet ambitieux à horizon long. La levée va financer ces recherches ». Sur les risques associés à ces nouveaux modèles, Yann le Cun estime dans les colonnes de [Libération le 10 mars 2026](#) que, « *à la fin, la décision de quelle est la meilleure utilisation de l'IA pour la société ne devrait pas être dans les mains de quelqu'un comme moi, ou comme mes collègues. C'est à la société et à ses institutions démocratiques de décider.* »