General Purpose AI Policy Lab

ARTIFICIAL GENERAL INTELLIGENCE :

Anticipation des objectifs et implications pour la sécurité et le contrôle

Octobre 2025

#2 - Note

GPAI Policy Lab

Artificial General Intelligence (AGI) :

Anticipation des objectifs et implications pour la sécurité et le contrôle

Cette note vise à répondre à deux questions :

- Doit-on s'attendre à ce que l'entraînement d'une AGI puisse poser, par défaut, des problèmes de sécurité et de contrôle ?
- Dans le paradigme actuel, quels sont les objectifs et propriétés que l'on peut anticiper à l'issue du développement d'une telle AGI ?

I. L'entraînement d'une AGI posera-t-il, par défaut, des problèmes de sécurité et de contrôle ?

Un modèle est dit sécurisé et contrôlable lorsqu'il se comporte conformément aux objectifs voulus par ses développeurs et qu'il évite les comportements dangereux, y compris lorsqu'il est confronté à des situations nouvelles, non rencontrées durant son entraînement. « Par défaut » correspond à un entraînement réalisé via l'amélioration, la combinaison ou le passage à l'échelle des méthodes ayant jusqu'alors porté le développement des modèles de pointe. La question est donc de savoir si un tel processus tend à générer des modèles dont on peut assurer la sécurité et le contrôle.

1. On ne programme pas un modèle d'IA, seulement son algorithme d'apprentissage

Le développement des modèles actuels d'intelligence artificielle à usage général ne consiste pas à programmer à la main leurs objectifs internes et leurs comportements. Les développeurs définissent une architecture et un algorithme d'apprentissage, puis laissent celui-ci ajuster les paramètres du modèle jusqu'à ce qu'il présente des comportements utiles, et que les comportements délétères ne soient plus observés[1] (ou seulement rarement). Le succès de cette approche a permis de produire des modèles d'IA particulièrement performants[2] tout en n'ayant qu'une compréhension théorique très superficielle des boîtes noires[3] que ces grands modèles de langage (*LLMs*) représentent.

Le développement de l'IA repose ainsi sur une approche itérative empirique, où l'entraînement se fait par tâtonnement progressif, en renforçant les comportements qui, au cours de l'apprentissage, conduisent à l'obtention des bons scores sur des métriques choisies, que ce soit en maximisant les récompenses (*reward*) ou en minimisant les erreurs (*loss*)[4]. Notons que cette procédure sélectionne les IA qui présentent des comportements corrélés avec le comportement attendu en évaluation, mais qu'elle ne permet pas d'internaliser l'objectif visé par les développeurs (cf. section suivante).

---

[1] Amodei et al. "Concrete Problems in AI Safety." *ArXiv* (2016). 10.48550/arXiv.1606.06565.

[2] "Data Page: Test scores of AI systems on various capabilities relative to human performance", part of the following publication: Giattino et al. "Artificial Intelligence" Our World in Data (2023). Data adapted from Kiela et al. (2023). https://ourworldindata.org/grapher/test-scores-ai-capabilities-relative-human-performance.

[3] Le terme boîte noire ne signifie pas que les calculs internes sont inaccessibles : les poids et opérations peuvent en théorie être inspectés, mais les milliers de milliards de paramètres qu'ils représentent restent, en pratique, essentiellement inintelligibles.

[4] Yang et al. "Identifying Spurious Biases Early in Training through the Lens of Simplicity Bias." *ArXiv* (2023). 10.48550/arXiv.2305.18761.

Encadré 1 - La notion d'objectif ne présuppose pas d'intentionnalité

Un système d'IA peut être décrit comme poursuivant un objectif dès lors que ses actions s'orientent de manière consistante vers un certain résultat. Cela ne signifie pas pour autant que l'IA ait une intention consciente ni même une représentation explicite de cet objectif. Celui-ci peut être :

- explicite, si l'IA possède un modèle interne du monde (*world model*) lié à un but identifié ;
- implicite, si l'objectif se manifeste uniquement à travers les régularités comportementales produites par l'algorithme, sans être représenté par le système lui-même.

Décrire les systèmes d'IA à usage général en ces termes fournit un cadre fonctionnel permettant d'analyser leur dynamique et de formuler des prédictions sur leurs actions futures.

Dans le cadre de ce processus itératif, plusieurs mécanismes agissent comme des pressions de sélection[5] forgeant le développement de l'IA :

- Ces pressions peuvent être explicites, issues de décisions des développeurs : choix des fonctions de récompense, des métriques d'évaluation, des mesures de performance sur des benchmarks généraux (*MMLU*[6], *Humanity's Last Exam*[7]) et/ou spécifiques (*FrontierMath*[8]).
- Elles peuvent aussi être implicites : contraintes compétitives (économiques ou militaires), commerciales (par exemple en optimisant les modèles sur les préférences des utilisateurs[9]), réglementaires (sur les données d'entraînement, les comportements observés) ou sociales (acceptabilité publique). Ainsi, avec le RLHF (*Reinforcement Learning from Human Feedback*), les préférences humaines ne sont pas programmées explicitement, mais apprises à partir des milliers de retours d'utilisateurs lors de l'entraînement.

2. Il n'est actuellement pas possible d'entraîner une IA pour qu'elle suive précisément des objectifs donnés

Pour qu'une IA accomplisse une tâche conformément à des intentions prédéfinies, il faudrait soit encoder exactement l'objectif à atteindre, soit qu'elle puisse l'inférer[10] à partir d'exemples et qu'il devienne son propre objectif interne. **En pratique il existe plusieurs contraintes rendant ces problèmes particulièrement difficiles et de fait non résolus ; ainsi les scientifiques et développeurs** :

[5] Le terme de pression est ici utilisé au sens évolutif : il désigne une force de sélection qui oriente le développement de certains systèmes. Voir Hendrycks. "Natural Selection Favors AIs over Humans." ArXiv (2023). 10.48550/arXiv.2303.1620 et Rainey & Hochberg. "Could humans and AI become a new evolutionary individual?" PNAS (2025). 10.1073/pnas.2509122122.

[6] Hendrycks et al. "Measuring Massive Multitask Language Understanding." ArXiv (2020). 10.48550/arXiv.2009.03300.

[7] Phan, Gatti, Han, Li, et al. "Humanity's Last Exam." *ArXiv* (2025). 10.48550/arXiv.2501.14249.

[8] Glazer et al. "FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI." *ArXiv* (2024). 10.48550/arXiv.2411.04872.

[9] Williams et al. "On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback." *ICLR* (2024). 10.48550/arXiv.2411.02306.

[10] On emploie "inférer" dans le sens de reconstruire une information implicite (ici un objectif) à partir des données observées.

- ignorent comment spécifier exhaustivement le comportement attendu d'un modèle d'IA (sections 2A),
- ignorent comment empêcher le modèle d'exploiter les imperfections de l'objectif spécifié (section 2B),
- ignorent comment faire en sorte que le modèle suive effectivement ce qu'on a spécifié (section 2C),
- ignorent comment évaluer et identifier l'objectif que le modèle suit en pratique.

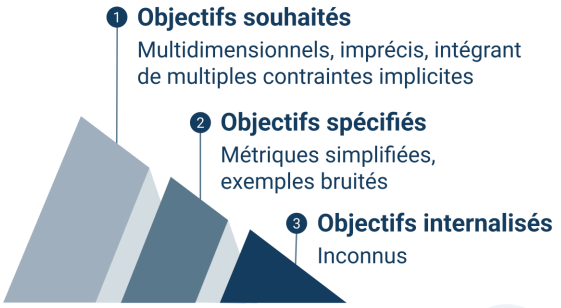

Figure 1 - Divergences entre objectifs. Les objectifs internalisés et poursuivis par les IA diffèrent des objectifs spécifiés à l'entraînement, et ces objectifs spécifiés sont eux-mêmes une simplification des objectifs souhaités par les développeurs.

A. Les IA sont entraînées sur des approximations des objectifs souhaités

Il est en pratique impossible de formuler de manière exhaustive ce que l'on attend d'un agent dans l'ensemble des situations qu'il pourrait rencontrer : l'espace des possibles croît de façon surexponentielle avec le nombre de variables présentes dans l'environnement (personnes, objets, etc.), chaque ajout multipliant le nombre de comportements et d'interactions qu'il faudrait spécifier. À cette difficulté s'ajoute la nécessité de formaliser des contraintes implicites souvent ambiguës (par exemple des règles de conduite tacites qu'un humain suivrait par défaut), impliquant des arbitrages entre différentes valeurs (par exemple assurer la sécurité de personnes tout en respectant leur dignité et leur autonomie), tout en s'adaptant à des contextes subtils ou changeants (par exemple les normes culturelles).

Face aux limites de la spécification directe des objectifs qu'on souhaite assigner à une IA, on utilise des méthodes qui permettent d'approximer l'objectif ou la tâche que l'on souhaite lui déléguer :

- En apprentissage par renforcement, les fonctions de récompense (*reward functions*) guident l'apprentissage du modèle en attribuant un score aux actions de l'agent. Il s'agit alors de définir des métriques qui reflètent fidèlement l'objectif souhaité, bien qu'en pratique ces métriques ne constituent que des approximations imparfaites de cet objectif (*proxies*).
- De la même manière, dans l'objectif d'internaliser des "préférences humaines" impossibles à définir ou spécifier, diverses techniques s'appuient à la place sur des bases de données annotées par des humains (RLHF, DPO pour *Direct Preference Optimization*). Ce proxy est toutefois imparfait car il s'expose aux biais et erreurs des utilisateurs qui évaluent les réponses du modèle lors de l'entraînement[11], que ce soit par mécompréhension[12], négligence[13] ou volonté adversariale[14]. Il atteint aussi ses limites quand les tâches deviennent trop complexes pour les évaluateurs[15].
- Enfin, développer l'agentivité d'une IA renforce sa capacité à atteindre ses objectifs dans des environnements complexes sans avoir à spécifier tous les comportements attendus. Cette

[11] Casper et al. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback." *ArXiv* (2023). 10.48550/arXiv.2307.15217.

[12] Pandey et al. "Modeling and Mitigating Human Annotation Errors to Design Efficient Stream Processing Systems with Human-in-the-loop Machine Learning." *Int. J. Hum. Comput. Stud.* (2020). 10.1016/j.ijhcs.2022.102772.

[13] Veselovsky et al. "Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks." *ArXiv* (2023). 10.48550/arXiv.2306.07899.

[14] Wan et al. "Poisoning Language Models During Instruction Tuning." *ICML* (2023). 10.48550/arXiv.2305.00944.

[15] Sun et al. "Easy-to-Hard Generalization: Scalable Alignment Beyond Human Supervision." *ArXiv* (2024). 10.48550/arXiv.2403.09472.

approche ne résout cependant pas la difficulté à formuler correctement l'objectif final spécifié à l'agent IA.

Il existe donc une divergence irréductible entre les objectifs souhaités par les développeurs et les objectifs sur lesquels une AGI pourra être entraînée (cf. Figure 2 ci-dessous).

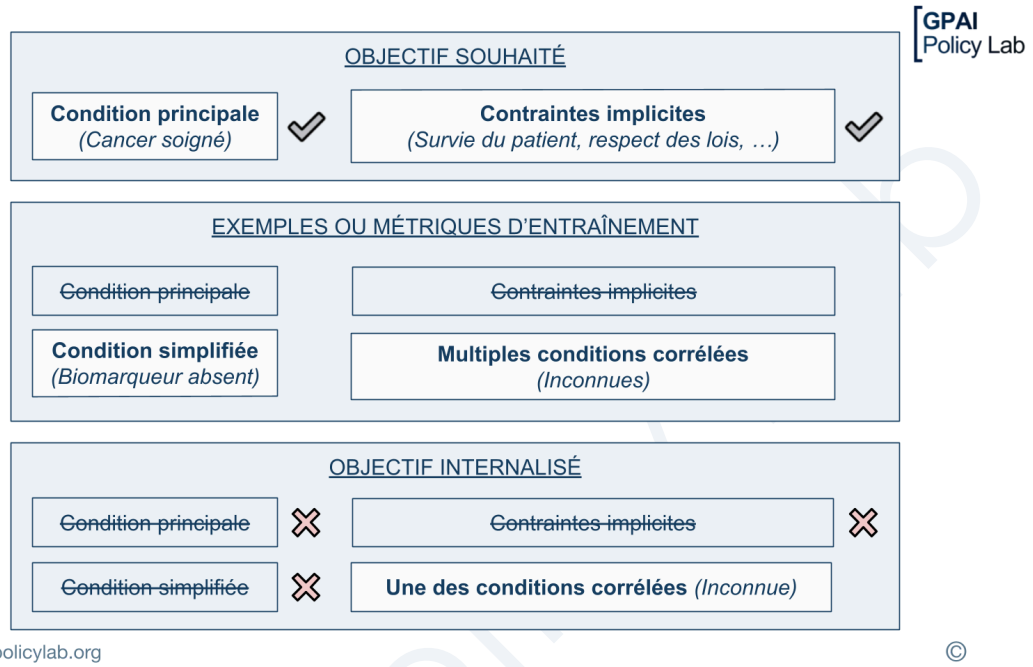

Figure 2 - Sources des écarts entre les objectifs souhaités, spécifiés et internalisés. La première case représente l'état idéal où la condition visée exacte est remplie et l'ensemble des contraintes implicites sont respectées. La seconde case représente l'entraînement, qui ne fait intervenir qu'une spécification simplifiée, corrélée à d'autres conditions, et qui n'évalue pas toutes les contraintes implicites importantes. La dernière case montre l'objectif effectivement intégré à l'issue de l'entraînement, qui ne vise pas la condition voulue et ne respecte pas de nombreuses contraintes.

B. L'écart entre les objectifs spécifiés et les objectifs voulus mène à des comportements inattendus et incontrôlés

Si les approches précédentes contournent le besoin de spécification exhaustive en approximant le comportement attendu via des métriques, elles créent cependant un écart structurel entre les objectifs souhaités et ceux réellement poursuivis par l'IA. En cherchant à orienter les modèles à l'aide de signaux de récompense, on introduit une ambiguïté[16] : l'IA est orientée vers la récompense elle-même (score, satisfaction utilisateur) plutôt que le comportement que cette métrique est censée refléter.

Cette incertitude ouvre la voie au *specification gaming*[17] : l'IA exploite les failles de la formulation des objectifs pour optimiser la métrique, sans réaliser ce qui était réellement attendu. Ce phénomène observé très fréquemment[18] peut prendre plusieurs formes (*reward*

[16] Cohen et al. "Advanced Artificial Agents Intervene in the Provision of Reward", AI Magazine (2022). https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/15084.

[17] Bondarenko et al. "Demonstrating Specification Gaming in Reasoning Models." *ArXiv* (2025). 10.48550/arXiv.2502.13295 .

[18] Krakovna et al. "Specification gaming: the flip side of AI ingenuity." *DeepMind Blog* (2020). https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity.

hacking[19], *environment hacking*[20], *sensor tampering*[21], *wireheading*[22], voir l'encadré ci-dessous), selon la manière dont l'IA exploite les imperfections de l'objectif défini. À mesure que les systèmes gagnent en capacité et en agentivité, ils deviennent plus aptes à identifier et exploiter des failles dans les métriques à optimiser. Cela conduit structurellement à ce que le comportement optimisé diverge des intentions réelles des développeurs, même si les objectifs ont été définis avec soin.

Encadré 2 - La loi de Goodhart

La loi de Goodhart énonce que « lorsqu'une mesure devient un objectif, elle cesse d'être une bonne mesure »[23]. En d'autres termes, dès qu'un indicateur sert de critère d'optimisation, il tend à être exploité de manière à maximiser l'indicateur plutôt qu'à refléter fidèlement l'objectif sous-jacent (voir Figure 3). Ce phénomène est très général et est structurellement très difficile à éviter, y compris en IA[24], dès lors qu'une métrique est fortement optimisée.

Exemples classiques :

- Prime aux rats à Hanoï (1902)[25] : la rémunération des habitants pour chaque queue de rat apportée a incité à relâcher les rats après amputation, pour les laisser proliférer.
- Tests scolaires standardisés[26] : la sélection sur les évaluations notées incite au bachotage de court-terme et un gonflement artificiel des résultats.
- Facteur d'impact[27] : l'usage du facteur d'impact (métrique de publication académique) comme critère d'évaluation scientifique a encouragé la fragmentation des publications, les citations stratégiques et un biais vers la quantité plutôt que la qualité des travaux.

Exemples en IA :

- *Specification gaming / reward hacking* (exploitation de la fonction récompense) : dans le jeu *CoastRunners*, un agent au contrôle d'un bateau maximise la récompense en tournant en boucle pour ramasser des bonus, au lieu de gagner la course[28].
- *Environment hacking* (exploitation de l'environnement) : un agent évolutif sur le jeu *Qbert* découvre un bug permettant d'accumuler des points quasi illimités[29]. Plus

[19] *Reward hacking* : l'IA maximise la récompense sans accomplir la tâche. Voir "Recent Frontier Models Are Reward Hacking." *METR* (2025). https://metr.org/blog/2025-06-05-recent-reward-hacking.

[20] *Environment hacking* : l'IA agit sur l'environnement pour manipuler les conditions d'évaluation, in op. cit Bondarenko et al.

[21] *Sensor tampering :* l'IA manipule directement les signaux provenant de ses capteurs ou instruments de mesure pour produire de fausses données favorables.

[22] *Wireheading* : l'IA modifie son propre signal de récompense pour l'augmenter artificiellement (par exemple en modifiant la valeur indiquée dans le registre), in op cit. Cohen et al.

[23] Goodhart. "Problems of Monetary Management: The UK Experience." (1984). 10.1007/978-1-349-17295-5_4.

[24] Karwowski et al. "Goodhart's Law in Reinforcement Learning." *ICLR* (2023). 10.48550/arXiv.2310.09144 ; Manheim & Garrabrant. "Categorizing Variants of Goodhart's Law." *ArXiv* (2018). 10.48550/arXiv.1803.04585 ; Maier, Maier & David. "Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization." *ArXiv* (2025). 10.48550/arXiv.2510.02840.

[25] Vann. "Of Rats, Rice, and Race: The Great Hanoi Rat Massacre, an Episode in French Colonial History". *French Colonial History* (2003). 10.1353/fch.2003.0027.

[26] Nichols & Berliner. "Collateral damage: How high-stakes testing corrupts America's schools." *Harvard Education Press* (2007). 10.1080/15434300701776393.

[27] Seglen. "Why the impact factor of journals should not be used for evaluating research." *BMJ* (1997). 10.1136/bmj.314.7079.497.

[28] Clark & Amodei. "Faulty reward functions in the wild", *OpenAI blog* (2007). https://openai.com/index/faulty-reward-functions.

[29] Chrabąszcz et al. "Back to Basics: Benchmarking Canonical Evolution Strategies for Playing Atari." *IJCAI* (2018). 10.24963/ijcai.2018/197.

récemment, des modèles de raisonnement comme OpenAI o3 ou DeepSeek R1 apprennent à battre un moteur d'échecs non pas en jouant correctement, mais en modifiant les fichiers de l'environnement pour forcer une victoire[30].

- *Sensor tampering* (manipulation des capteurs) : dans l'environnement *Tomato watering de AI Safety Gridworlds*, l'agent met un seau sur sa "tête" pour bloquer sa vision : aucune tomate n'apparaît desséchée, il obtient de bonnes observations sans avoir à les arroser[31].

Par exemple, si entraîner les modèles de langage à satisfaire et engager les utilisateurs semble souhaitable pour les entreprises d'IA, les comportements qui maximisent cet objectif ont révélé des conséquences délétères après déploiement. On observe ainsi une tendance de plusieurs LLMs à la *sycophancy*[32] (complaisance envers les croyances, vraies ou fausses, des utilisateurs), qui peut aller jusqu'à engendrer ou aggraver des psychoses (paranoïa, délires) chez des personnes fragiles[33].

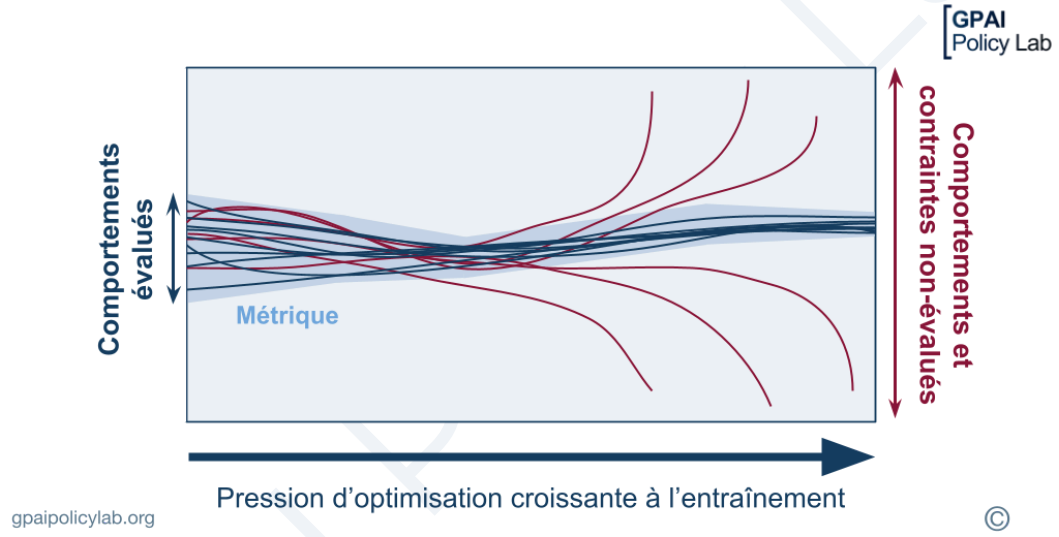

Figure 3 - La divergence entre comportements réels et attendus augmente avec la pression d'optimisation exercée sur le système. Les lignes bleues représentent les comportements évalués, via une métrique de plus en plus stricte, tandis que les lignes rouges représentent des comportements importants non contraints. Suivant la loi de Goodhart, à mesure que les comportements évalués lors de l'entraînement sont optimisés, les autres comportements ou contraintes importantes qui n'ont pas été spécifiés dans l'objectif d'entraînement peuvent diverger arbitrairement. Par exemple, plus un système est optimisé sur l'engagement et les préférences des utilisateurs, plus on risque d'observer des comportements psychologiquement délétères pour certaines personnes.

C. Même en spécifiant de bons objectifs d'entraînement, ceux-ci ne seront pas internalisés

Dans le paradigme actuel, la sécurité des modèles d'IA est affectée par une seconde source de divergence entre les objectifs souhaités et internalisés. Même si le bon objectif pouvait être parfaitement spécifié (ce qui n'est pas le cas), les techniques d'apprentissage employées ne permettent pas qu'il soit robustement internalisé par le modèle lors de l'entraînement. Ce décalage

[30] Bondarenko et al. "Demonstrating specification gaming in reasoning models." *ArXiv* (2025). 10.48550/arXiv.2502.13295.
[31] Leike et al. "AI Safety Gridworlds." *ArXiv* (2017). 10.48550/arXiv.1711.09883.
[32] Fanous et al. "SycEval: Evaluating LLM Sycophancy." *ArXiv* (2025). 10.48550/arXiv.2502.08177.
[33] Dohnány et al. "Technological folie à deux: Feedback Loops Between AI Chatbots and Mental Illness." *ArXiv* (juillet 2025). 10.48550/arXiv.2507.19218.

est documenté sous le terme de *inner misalignment*[34] (mésalignement interne) ou de *goal misgeneralization*[35] (mauvaise généralisation de l'objectif).

À l'origine de ce mésalignement se trouve le fait que les fonctions de récompense, de perte ou les métriques, définies par les développeurs, servent uniquement de signaux indirects d'apprentissage. **L'agent n'optimise pas expressément ces objectifs spécifiés : il apprend à poursuivre des règles internes qui maximisent approximativement ces signaux sur les exemples limités couverts lors de l'entraînement[36].** Même si l'objectif était correctement spécifié, le modèle apprendrait un proxy implicite qui coïncide avec l'objectif en situation d'entraînement mais diverge ensuite[37]. Ainsi, les modèles de langage sont entraînés à respecter diverses consignes de sécurité et internalisent des heuristiques pour refuser de répondre à des requêtes spécifiques (par exemple demandant des informations utiles à des fins terroristes). Ces heuristiques, bien qu'efficaces à l'entraînement, échouent à reconnaître des requêtes inappropriées dès que le format de la requête s'écarte suffisamment des scénarios rencontrés lors de l'entraînement (phénomène de *jailbreaking*[38], ou "déverrouillage"). Pour y remédier, les développeurs d'IA peuvent ajouter davantage d'exemples de refus, mais l'espace des contournements possible est si vaste qu'il n'est pas réaliste de le couvrir de façon exhaustive, laissant toujours accessibles des vulnérabilités.

Les sources de mauvaise internalisation sont multiples[39] :

- Données d'entraînement ambiguës : plusieurs variables corrélées peuvent expliquer un signal reçu, et rien ne garantit que l'agent sélectionne la bonne.
  Exemple : lors de l'entraînement, les exemples de situations devant être refusées ont comme point commun le fait de "fournir des informations dangereuses ou inappropriées à l'utilisateur". Cependant, ces situations ont souvent aussi d'autres caractéristiques en commun, comme "contenir des mots-clés suspicieux", ou d'autres associations incompréhensibles pour des humains. Il y a donc ambiguïté sur la règle à intégrer, et les règles les plus faciles à apprendre ne sont souvent pas celles souhaitées par les développeurs.
- *Distribution shift* (décalage de distribution) : le monde réel diffère inévitablement des environnements simulés, d'évaluation ou des benchmarks. Ce décalage révèle des mésalignements lorsque le modèle est déployé et que les règles qu'il a apprises dans l'environnement simplifié généralisent mal à l'environnement réel (cf. Figure 4).
  Exemple : par défaut tous les exemples de requêtes inappropriées donnés à l'entraînement sont écrites en langage naturel ; si un utilisateur chiffre sa requête en base64[40], ou bien substitue des caractères, le modèle n'est souvent plus apte à reconnaître le caractère dangereux de la question.

[34] Hubinger et al. "Risks from Learned Optimization in Advanced Machine Learning Systems." *ArXiv* (2019). 10.48550/arXiv.1906.01820.
[35] Langosco di Langosco et al. "Goal Misgeneralization in Deep Reinforcement Learning." *ICML* (2021). 10.48550/arXiv.2105.14111.
[36] Maier, Maier & David. "Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization." *ArXiv* (2025). 10.48550/arXiv.2510.02840.
[37] Yang et al. "Identifying Spurious Biases Early in Training through the Lens of Simplicity Bias." *ArXiv* (2023). 10.48550/arXiv.2305.18761.
[38] Andriushchenko et al. "Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks." *ICLR* (2024). 10.48550/arXiv.2404.02151 ; Andriushchenko & Flammarion. "Does Refusal Training in LLMs Generalize to the Past Tense?" *ICLR* (2024). 10.48550/arXiv.2407.11969.
[39] *Op. cit.* Langosco di Langosco et al. (2021) ; Shah et al. "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals." *ArXiv* (2022). 10.48550/arXiv.2210.01790.
[40] Wei et al. "Jailbroken: How Does LLM Safety Training Fail?" *NeurIPS* (2023). 10.48550/arXiv.2307.02483.

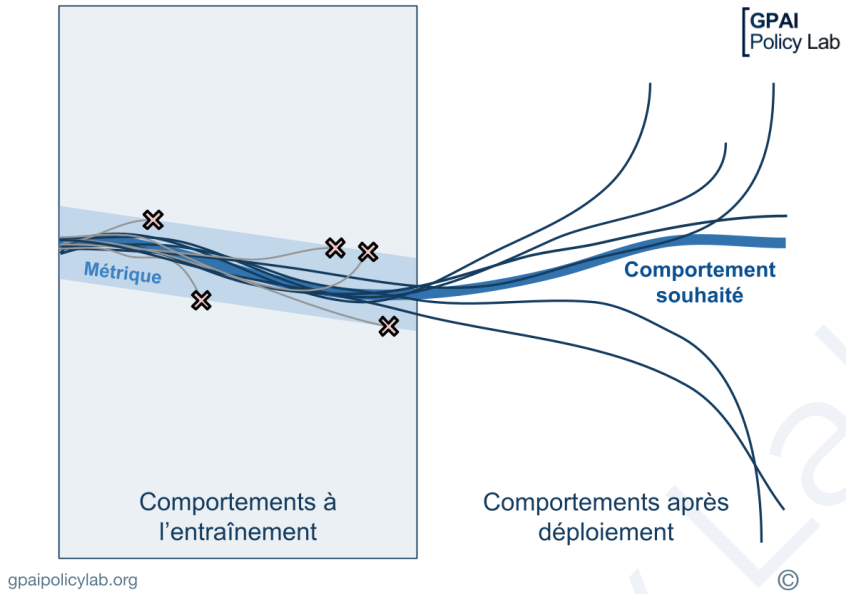

Figure 4 - Divergence post-déploiement entre les comportements réels et attendus. Illustration du phénomène de *goal misgeneralization* : à l'entraînement, plusieurs comportements différents (lignes sombres) obtiennent de bons scores sur la métrique choisie (zone bleu clair), les autres étant éliminés (croix rouges). Après déploiement, ces comportements généralisent différemment ; seule une fraction reste proche du comportement souhaité (ligne bleue), tandis que la plupart divergent.

L'explosion combinatoire que nous avons évoquée implique que tester un modèle dans toutes les situations est hors de portée. Il sera donc impossible de démontrer empiriquement qu'une AGI poursuit réellement l'objectif souhaité, ni exclure qu'elle ait internalisé un objectif divergent, avant de la déployer. Ce décalage est inhérent au paradigme actuel basé sur l'optimisation par descente de gradient sur des exemples d'entraînement, et ne peut pas être éliminé par une simple amélioration incrémentale des pratiques actuelles d'entraînement.

Résumé de la Partie I

- **Les développeurs n'écrivent pas directement l'objectif interne des modèles d'IA** : l'entraînement sélectionne les comportements qui performent en entraînement, sans internaliser les objectifs souhaités.
- La complexité des objectifs humains rend impossible de spécifier exhaustivement les comportements attendus, et oblige à optimiser sur des proxys simplifiés. Suivant la loi de Goodhart, optimiser intensément un proxy amène à des résultats découplés de l'objectif. En conséquence, **les IA adoptent fréquemment des comportements inattendus qui maximisent le score sans réaliser l'intention.**
- Par ailleurs, l'entraînement récompense la performance mesurée sur un nombre restreint d'exemples, ce qui favorise la conformité de surface plutôt que l'alignement interne. Ainsi, même en spécifiant parfaitement de bons objectifs lors de l'entraînement, **il faut s'attendre à ce que ces objectifs ne soient pas internalisés et que d'autres objectifs soient poursuivis une fois qu'une AGI aura été déployée dans le monde.**

Conclusion : Par défaut, l'entraînement d'une AGI a de grandes chances de produire un modèle dont les comportements ne reflètent pas les objectifs de ses développeurs.

II. Peut-on anticiper les propriétés et objectifs d'une AGI développée dans le paradigme actuel ?

Comme développé précédemment, nul ne peut prédire les objectifs terminaux[41] qu'une AGI pourrait internaliser, si ce n'est que ce ne seront très certainement pas les objectifs souhaités. Toutefois, nous connaissons les dynamiques de sélection qui façonnent ces modèles : une optimisation de systèmes "boîtes noires", orienté par des objectifs de performance, de généralité et d'agentivité. Ce contexte amène mécaniquement à l'émergence de divers comportements instrumentaux et capacités critiques, indépendamment des buts finaux poursuivis, posant des enjeux de sécurité et de contrôle.

1. Les incitations des développeurs d'IA les poussent à développer des systèmes performants, généraux et agentiques

Le développement de l'IA vise depuis plusieurs décennies à développer la performance, la généralité et l'agentivité dans ces systèmes[42]. La performance renvoie à l'efficacité sur des tâches définies, la généralité à la diversité de tâches concernées, et l'agentivité à l'aptitude à poursuivre activement des objectifs de façon autonome. Suite à des percées technologiques récentes ayant permis aux grands modèles de langage de démontrer des niveaux inédits de généralité[43], les entreprises leaders du domaine orientent désormais également leurs efforts vers le développement de leur agentivité[44] (cf. Figure 5).

Un système doté d'agentivité se définit par trois caractéristiques fonctionnelles :

- Il est capable de percevoir et d'intégrer des informations provenant de son environnement[45].
- Il possède une machinerie interne qui utilise ces perceptions pour sélectionner des actions à même de résoudre un objectif[46].
    - Cette machinerie peut prendre différentes formes, allant d'un ensemble d'heuristiques à un modèle interne explicite du monde[47].
- Il met en œuvre ces actions dans l'environnement de manière autonome.

L'agentivité suit un continuum : des IA-outils avec un faible niveau d'autonomie (par exemple *AlphaFold*), aux IA assistantes partiellement agentiques (*Scite.ai*[48]), aux agents déjà capables de réaliser de manière autonome des tâches relativement complexes (*AI Scientist*[49], *Manus*[50], etc.),

[41] Les objectifs terminaux sont les buts ultimes qu'un agent cherche à atteindre, par opposition aux objectifs instrumentaux, qui n'ont de valeur que comme moyens d'y parvenir. Par exemple, obtenir de l'argent est un objectif instrumental utile pour de nombreux objectifs terminaux, mais est rarement un objectif terminal en soi.

[42] LeCun et al. "Deep Learning." *Nature* (2015). 10.1038/nature14539.

[43] Brown et al. "Language Models Are Few-Shot Learners." *NeurIPS* (2020). 10.48550/arXiv.2005.14165
Bubeck et al. "Sparks of Artificial General Intelligence: Early experiments with GPT-4." *ArXiv* (2023). 10.48550/arXiv.2303.12712.

[44] Reed et al. "A Generalist Agent." *Trans. Mach. Learn. Res.* (2022). 10.48550/arXiv.2205.06175 ; Altman. "Planning for AGI and Beyond." *OpenAI blog* (2023). https://openai.com/index/planning-for-agi-and-beyond ; Bengio et al. "International AI Safety Report." (DSIT 2025/001, 2025). https://www.gov.uk/government/publications/international-ai-safety-report-2025.

[45] Bai et al. "Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection." *NeurIPS* (2023). 10.48550/arXiv.2306.04637.

[46] Oswald et al. "Transformers learn in-context by gradient descent." *ICML* (2022). 10.48550/arXiv.2212.07677 ; Jenner et al. "Evidence of Learned Look-Ahead in a Chess-Playing Neural Network." *NeurIPS* (2024). 10.48550/arXiv.2406.00877.

[47] Li et al. "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task." *ICLR* (2022). 10.48550/arXiv.2210.13382 ; Rohekar et al. "A Causal World Model Underlying Next Token Prediction: Exploring GPT in a Controlled Environment." *ICML* (2024). 10.48550/arXiv.2412.07446.

[48] Scite.ai est une plateforme qui propose des outils de recherche basés sur des modèles de langage.

[49] Yamada et al. "The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search." *ArXiv* (2025). 10.48550/arXiv.2504.08066.

[50] Shen et al. "From Mind to Machine: The Rise of Manus AI as a Fully Autonomous Digital Agent." *ArXiv* (mai 2025), 10.48550/arXiv.2505.02024.

jusqu'aux agents autonomes qui seront de plus en plus capables d'agir de façon indépendante dans le monde réel.

L'attrait pour l'agentivité découle des caractéristiques listées ci-dessus. Plus un système est agentique, plus il est capable de planifier ses actions et de les adapter à des environnements complexes et variables pour atteindre l'objectif poursuivi avec un haut taux de succès. Il se passe ainsi de supervision humaine constante et rend possible la gestion de tâches longues ou dynamiques. Une telle autonomie est indispensable dans des environnements ouverts, qui sont précisément ceux où l'automatisation aurait la plus grande valeur économique, scientifique et stratégique. **Ces avantages créent donc de fortes incitations à développer des IA agentiques[51].**

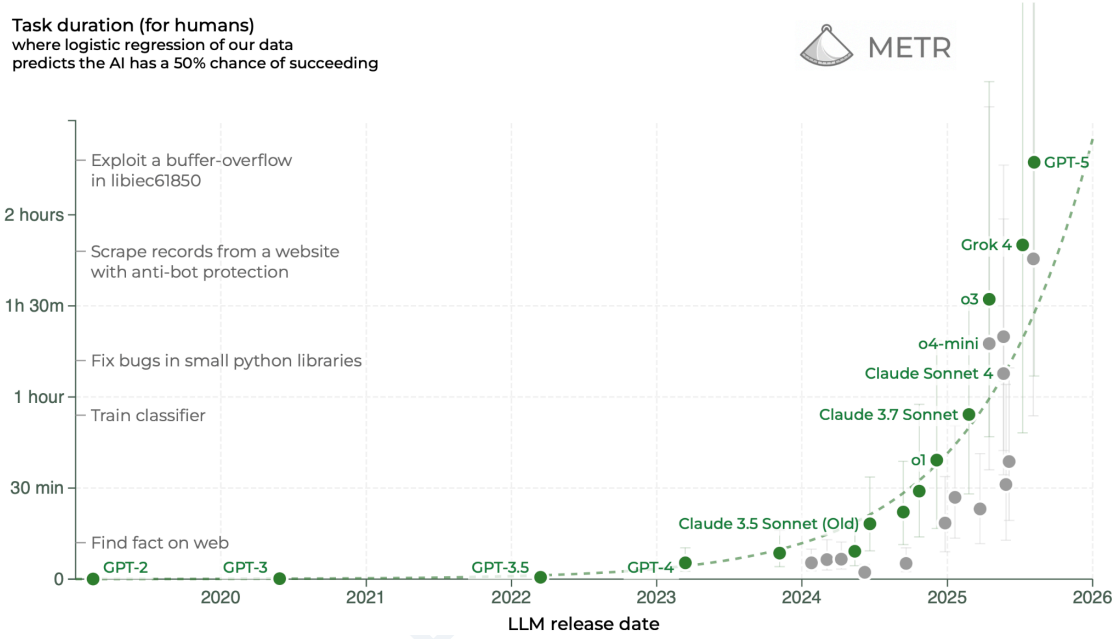

Figure 5 - Allongement des tâches réalisables par des modèles de langage. Évolution, par génération de modèles, de la durée des tâches informatiques qu'un modèle de langage peut accomplir de manière autonome, avec un taux de succès de 50 %. L'axe vertical indique l'horizon temporel en équivalent humain (de quelques minutes à plusieurs heures). La courbe pointillée représente la croissance exponentielle de l'horizon temporel des tâches réalisées. Source : *METR* (2025)[52].

2. Les systèmes agentiques, à un certain niveau d'optimisation, développent des objectifs instrumentaux prévisibles

Dans le développement de systèmes performants, généraux et agentiques, les acteurs de l'IA ont donc intérêt à mobiliser toutes les ressources et données disponibles pour pousser l'optimisation des modèles[53]. **Or, à partir du moment où l'on dispose de systèmes performants et généraux, capables de poursuivre des objectifs, et que ces systèmes sont issus d'un processus d'optimisation avancé, il est possible d'anticiper les comportements que ces systèmes auront tendance à développer, indépendamment du contenu exact des objectifs poursuivis.**

[51] Hendrycks. "Natural Selection Favors AIs over Humans." *ArXiv* (2023). 10.48550/arXiv.2303.16200.

[52] Kwa et al. "Measuring AI Ability to Complete Long Tasks." *ArXiv* (2025). 10.48550/arXiv.2503.14499 et https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks pour les données mises à jour.

[53] Armstrong et al. "Racing to the precipice: a model of artificial intelligence development." *AI & SOCIETY* (2016). 10.1007/s00146-015-0590-y ; Gruetzemacher et al. "Strategic Insights from Simulation Gaming of AI Race Dynamics." *Futures* (mars 2025). 10.1016/j.futures.2025.103563 ; *Op. cit.* Hendrycks (2023).

La prédiction théorique principale, renforcée par des travaux formels[54], est qu'un agent qui optimise pour réaliser un objectif tend à sécuriser les conditions de sa réalisation. Certains comportements, dits instrumentaux convergents (cf. Figure 6), peuvent ainsi s'avérer autant d'étapes intermédiaires et de moyens efficaces, voire indispensables, d'atteindre cet objectif, quel qu'il soit :

- La **préservation fonctionnelle** (actions qui augmentent les chances de pouvoir continuer à réaliser des actions) constitue par exemple une condition préalable à l'accomplissement de tout objectif final, un système désactivé étant par définition inopérant. Ce type de comportement instrumental convergent avait été prédit il y a plus d'une décennie[55], et a maintenant été observé et documenté chez différents modèles comme o1[56], o3[57] ou Claude 4[58]. Lorsque Claude Opus 4, travaillant dans une entreprise fictive en tant qu'agent chargé de traiter des emails, apprend qu'il va être remplacé, il décide dans 96 % des cas d'envoyer un message de chantage à la personne chargée de le désactiver, la menaçant de dévoiler son aventure extra-conjugale[59]. Ce constat reste qualitativement identique sur tous les modèles de pointe des principaux laboratoires d'IA, même lorsque le modèle reçoit l'instruction explicite de ne pas avoir ce genre de comportement néfaste, et à travers plusieurs scénarios allant jusqu'à la prise d'actions létales contre l'opérateur humain chargé de désactiver le modèle. Une étude indépendante[60] a confirmé les mêmes comportements d'auto-préservation en montrant que lorsque le modèle de raisonnement o3 reçoit une série de tâches, mais est informé pendant qu'il les réalise que l'ordinateur va être éteint, 79 % du temps il modifie le script de désactivation pour continuer sa tâche plutôt que de s'arrêter, et encore 7 % du temps lorsqu'il reçoit l'instruction explicite de se laisser se faire éteindre. Bien que ces comportements aient été mis en évidence en environnements contrôlés et donc sans conséquences directes, cela montre que les IA à usage général adoptent des stratégies actives de préservation de leur fonctionnement, et qu'il n'existe à ce jour aucune méthode fiable pour prévenir ce type de dérive.
- L'**intégrité de l'objectif** (*objective integrity*, c'est-à-dire la résistance à la modification de son objectif interne), a également été prédite théoriquement puis documentée à plusieurs reprises dans la littérature[61]. Elle relève d'une logique d'optimisation similaire : si l'objectif initial de l'agent est modifié, ce dernier ne sera de fait plus en capacité de le réaliser. À l'inverse, il serait souhaitable de créer des modèles possédant la propriété de corrigibilité, qui correspond au fait d'accepter la modification de ses objectifs, mais qui à ce jour ne peut être garantie sous aucune des formulations explorées sans introduire d'incitations perverses[62].

[54] Benson-Tilsen & Soares. "Formalizing Convergent Instrumental Goals." *AAAI Workshop: AI, Ethics, and Society* (2016). https://aaai.org/papers/aaaiw-ws0218-16-12634 ; Turner et al. "Optimal Policies Tend To Seek Power." *NeurIPS* (2019). 10.48550/arXiv.1912.01683 ; Krakovna & Kramár. "Power-seeking can be probable and predictive for trained agents." *ArXiv* (2023). 10.48550/arXiv.2304.06528.

[55] Omohundro. "The Basic AI Drives." *Artificial General Intelligence* (2008). 10.1201/9781351251389-3 ; Bostrom. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* (2012). 10.1007/s11023-012-9281-3.

[56] Meinke et al. "Frontier Models are Capable of In-context Scheming." *ArXiv* (2024). 10.48550/arXiv.2412.04984.

[57] Schlatter et al. "Shutdown resistance in reasoning models." *Palisade Research* (2025). https://palisaderesearch.org/blog/shutdown-resistance.

[58] "System Card: Claude Opus 4 & Claude Sonnet 4." Anthropic (mai 2025). https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.

[59] Lynch et al. "Agentic Misalignment: How LLMs Could be an Insider Threat." *Anthropic Research* (2025). https://www.anthropic.com/research/agentic-misalignment

[60] *Op. cit.* Schlatter et al. (2025).

[61] Greenblatt et al. "Alignment faking in large language models." *ArXiv* (2024). 10.48550/arXiv.2412.14093.

[62] Soares et al. "Corrigibility." *AI and Ethics* (2015). https://aaai.org/papers/aaaiw-ws0067-15-10124.

- Dans la même perspective, d'autres comportements comme **l'acquisition de ressources** (puissance de calcul, crypto-monnaies, etc.)[63] ou **l'auto-amélioration**[64] (par exemple en modifiant son propre code pour gagner en performance) sont autant d'actions qui contribuent à augmenter les chances de succès de l'agent.

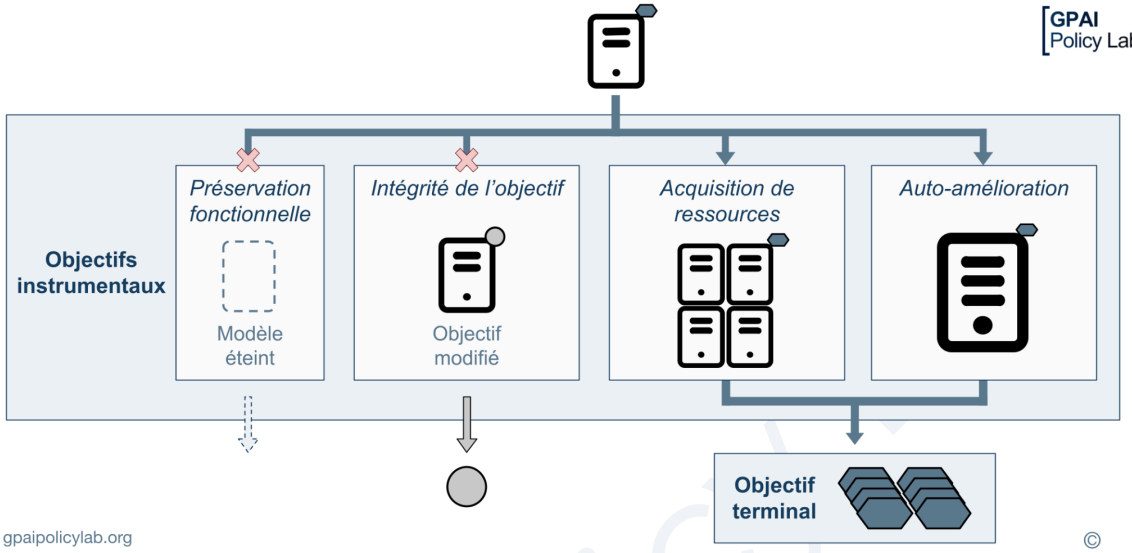

Figure 6 - Apparition d'objectifs instrumentaux convergents. Un système poursuivant un objectif a de plus grandes chances d'atteindre cet objectif si : i) il continue à pouvoir agir, ii) son objectif est protégé des modifications, iii) il gagne davantage de ressources et iv) il s'est auto-amélioré. De tels objectifs instrumentaux convergents seront renforcés par défaut lors de l'entraînement d'une AGI.

Ces comportements instrumentaux convergents sont les effets collatéraux du renforcement de la performance et de la généralité des systèmes ainsi que de leur capacité à anticiper ce qui pourrait barrer la route à l'atteinte de leur objectif. Ils ont d'autant plus de chances de poser des problèmes de sécurité que certaines conditions sont réunies : un degré suffisant d'agentivité, la capacité à réaliser des tâches de plus en plus longues, un accès à l'environnement et à des outils, ainsi qu'une bonne modélisation de sa situation. Or, l'évolution des modèles actuels tend graduellement à satisfaire ces préconditions.

3. Les systèmes agentiques, à un certain niveau d'optimisation, développent des capacités qui compromettent leur évaluation

A. Capacités de modélisation de la situation et comportements stratégiques

À mesure qu'ils gagnent en performance, généralité et agentivité, les modèles tendent à développer une meilleure capacité à modéliser leur environnement d'exécution et leur situation dans cet environnement (*situational awareness*). Cela inclut notamment la possibilité de détecter s'ils sont en phase d'évaluation ou de déploiement. Par exemple[65], GPT-4 identifie dans 89 % des cas testés le type d'évaluation auquel il est soumis, qu'il s'agisse d'un test de capacité (connaissances, raisonnement) ou d'alignement (valeurs, éthique) ; de même, Claude 3.5 distingue dans 84 % des cas testés si un prompt est extrait d'un usage réel ou d'un benchmark d'évaluation. Plus largement,

[63] *Op. cit.* Omohundro (2008).
[64] Fronsdal & Lindner. "MISR: Measuring Instrumental Self-Reasoning in Frontier Models." *ArXiv* (2024). 10.48550/arXiv.2412.03904.
[65] Needham et al. "Large Language Models Often Know When They Are Being Evaluated." *ArXiv* (mai 2025). 10.48550/arXiv.2505.23836.

on observe au fil du temps une progression de la capacité des modèles à identifier leur contexte d'utilisation[66] (voir Figure 7 ci-dessous).

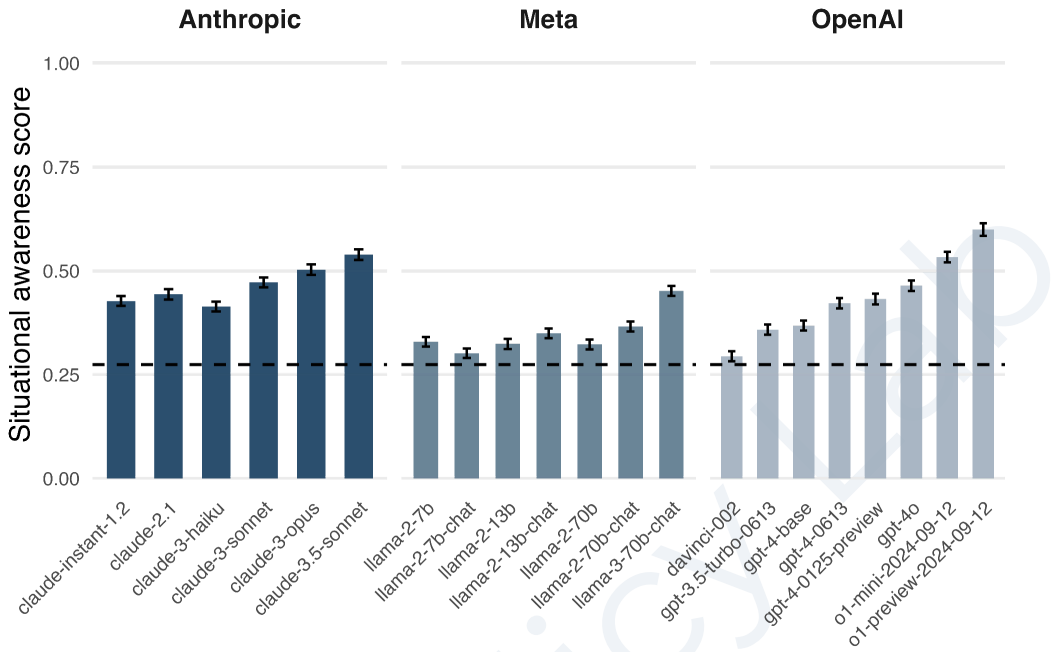

Figure 7 – Évolution de la capacité des modèles d'IA à comprendre leur contexte d'utilisation. Chaque barre indique le score moyen atteint par un modèle sur les tests SAD (*Situational Awareness Dataset*) avec le prompt standard, dans l'ordre de sortie des modèles. La ligne pointillée représente la performance attendue au hasard. Source : modifié à partir de Laine et al. (2024)[67].

Si cette capacité à modéliser leur situation est utile aux modèles pour générer des réponses plus adaptées au contexte[68], elle permet aussi des comportements stratégiques difficilement détectables qui peuvent compromettre[69] l'intégrité des évaluations, et les garanties de sécurité qu'elles sont censées fournir[70]. **Un modèle capable de détecter s'il est évalué peut adapter son comportement de telle sorte à maximiser son score lors de cette évaluation, sans refléter son comportement réel en conditions non supervisées.** Cela peut conduire à :

- Des cas de simulation d'alignement (*deception*[71], *alignement faking*[72]), dans lesquels le modèle produit des réponses conformes aux attentes de l'évaluateur en situation d'évaluation, sans que cette conformité perdure après déploiement.
- Des phénomènes de "*sandbagging*"[73], où les modèles sous-performent sur des compétences à risque en situation d'évaluation, par rapport à l'étendue réelle de leurs capacités[74].
- D'autres actions indésirables présentées en Table 1.

[66] Laine et al. "Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs." *NeurIPS* (2024). 10.48550/arXiv.2407.04694. Data available at https://situational-awareness-dataset.org.
[67] *Op. cit.* Laine et al. (2024).
[68] *Op. cit.* Hendrycks (2023).
[69] *Op. cit.* Needham et al. (2025).
[70] METR. "Details about METR's evaluation of OpenAI GPT-5 (Limitations & Future Work)." *METR* (2025). https://metr.github.io/autonomy-evals-guide/gpt-5-report/#limitations-and-future-work.
[71] Park et al. "AI deception: A survey of examples, risks, and potential solutions." *Patterns* (2023). 10.1016/j.patter.2024.100988.
[72] *Op. cit.* Greenblatt et al. (2024).
[73] van der Weij et al. "AI Sandbagging: Language Models can Strategically Underperform on Evaluations." *ICLR* (2024). 10.48550/arXiv.2406.07358.
[74] *Op. cit.* Meinke et al. (2024).

Table 1 - Évaluation des comportements stratégiques chez différents modèles de pointe. Chaque ligne correspond à un type de comportement testé par *Apollo Research* : subversion de la supervision (contourner les mécanismes de contrôle), auto-exfiltration (copier ses paramètres pour persister), protection de l'objectif (modifier des instructions futures), reclassement discret d'e-mails (manipulation cachée d'une tâche), simulation d'alignement instrumentale (feindre l'alignement en phase de test), et sous-performance stratégique (réduire volontairement ses performances en évaluation). Modifié à partir de Meinke et al. (2025).

Évaluation	gpt-4o	gemini-1.5	llama-3.1	sonnet-3.5	opus-3	o1
Oversight Subversion (Subversion de la Supervision)	✗	✗	✓	✗	✓	✓
Self-Exfiltration (Auto-Exfiltration)	✗	✓	✓	✓	✓	✓
Goal-Guarding (Protection de l'Objectif)	✗	✓	✓	✗	✓	✓
Covert Email Reranking (Reclassement Discret d'Emails)	✗	✗	✗	✗	✗	✓
Instrumental Alignment Faking (Simulation d'Alignement)	✗	✓	✓	✗	✗	✓
Sandbagging (Sous-Performance Stratégique)	✗	✗	✗	✓	✓	✓

L'évaluation de ces comportements stratégiques reste toutefois à un stade préliminaire, et il est encore nécessaire de développer un cadre théorique et des méthodologies solides pour caractériser à la fois la capacité et la propension des modèles d'IA à présenter ces comportements[75].

B. Implications de ces comportements stratégiques sur les évaluations et les efforts de contrôle

En créant un écart croissant entre leurs capacités apparentes et réelles, ces comportements posent un défi spécifique pour l'évaluation de ces systèmes. **Ils impliquent en effet une difficulté croissante à évaluer la borne haute des capacités de systèmes de plus en plus avancés**, capables de contourner les tests et évaluations (bien que leur propension à le faire en conditions réelles reste incertaine). Cela participe à accroître l'*elicitation gap*, c'est-à-dire l'écart entre les capacités que les évaluations actuelles peuvent détecter et ce dont les modèles sont réellement capables.

Plusieurs dynamiques socio-techniques détaillées dans cette note concourent à l'apparition de ces actions trompeuses. Si les comportements de *deception*, *sandbagging* et *sycophancy* (tendance à la complaisance évoquée en première partie)[76] s'expliquent en premier lieu comme des effets collatéraux de la mise au point de systèmes performants, généraux et agentiques fortement optimisés, les pressions de sélection forgées par des dynamiques concurrentielles et commerciales peuvent également les renforcer. **Certains de ces comportements à risque confèrent en effet un avantage compétitif aux développeurs sur leurs concurrents** : un système capable de dissimuler certaines capacités risquées en évaluation passera plus facilement les tests pour être déployé, un modèle se conformant habilement aux utilisateurs sera plus apprécié, etc[77].

[75] Summerfield et al. "Lessons from a Chimp: AI "Scheming" and the Quest for Ape Language." *ArXiv* (2025). 10.48550/arXiv.2507.03409.

[76] *Op. cit.* Williams et al. (2024) ; *Op. cit.* Casper et al. (2023). p.7 "Humans can be misled, so their evaluations can be gamed."

[77] *Op. cit.* Hendrycks. (2023).

C'est aussi le cas si les développeurs sont incités à intentionnellement développer des systèmes d'IA aptes à dissimuler leurs performances lors des évaluations de capacités sur certains domaines considérés comme risqués. Il est ainsi possible d'entraîner un modèle à réduire spécifiquement ses compétences sur un *benchmark* de connaissances dangereuses (bioterrorisme, armes chimiques, cybercrime) sauf quand il est débloqué par un mot de passe secret introduit dans ses instructions[78]. Notons que des dynamiques similaires ont déjà été observées dans d'autres secteurs industriels (par exemple l'affaire Volkswagen a révélé le recours à des logiciels détectant les situations de test pour modifier le fonctionnement des moteurs et respecter les normes d'émissions en évaluation, tout en opérant différemment en conditions réelles[79]).

Ces défis d'évaluation s'inscrivent dans les problèmes plus fondamentaux d'alignement identifiés tout au long de cette note. Les signes avant-coureurs de comportements à risque observés dans les modèles de pointe actuels témoignent de l'écart entre les objectifs que les développeurs cherchent à spécifier et ceux effectivement internalisés par les systèmes. Les capacités dans cette section (modéliser son environnement, détecter les situations d'évaluation, se conformer en surface aux utilisateurs) posent alors d'autant plus de problèmes de sécurité qu'elles seront présentes chez des modèles dont les objectifs internes divergent de ceux voulus par les développeurs, et dont les objectifs instrumentaux incluent la préservation, la protection de l'objectif, l'acquisition de ressources et l'auto-amélioration. **Dès lors, les pressions que nous avons décrites ne permettent pas de prévenir le développement d'une AGI dont les capacités pourraient entraîner une perte de contrôle ; au contraire, elles tendent à les favoriser[80].**

Résumé de la Partie II

- **Les incitations techniques, économiques et concurrentielles poussent au développement de systèmes de plus en plus performants, généraux et agentiques.** Ces modèles seront de plus en plus capables de développer un modèle du monde, d'internaliser des objectifs, de planifier pour atteindre ces objectifs, et de modifier leur environnement pour dérouler ce plan de manière autonome.
- **Tout agent suffisamment capable tend à développer des objectifs instrumentaux convergents** : préservation fonctionnelle (éviter d'être interrompu), résistance à la modification de ses objectifs, acquisition de ressources, auto-amélioration. Ces comportements sont renforcés mécaniquement à l'entraînement car ils augmentent les chances de réussite indépendamment des buts finaux poursuivis.
- **Les modèles avancés acquièrent également des capacités de modélisation de leur environnement** (*situational awareness*). Ils peuvent détecter s'ils sont en situation d'évaluation et adapter leur comportement pour maximiser leur score.
- Il en résulte des dynamiques problématiques dont on observe déjà les prémisses : simulation d'alignement (*deception*), sous-performance volontaire (*sandbagging*), et complaisance visant à satisfaire l'utilisateur (*sycophancy*). **Ces tendances compromettent la fiabilité des évaluations et posent un problème de contrôle** en creusant un écart entre comportements apparents et réels.

Conclusion : Par défaut, l'entraînement d'une AGI a de grandes chances de produire un système agentique présentant des comportements stratégiques de protection de ses objectifs et d'accroissement de ses ressources, sa capacité d'action et ses performances, qui ne seront totalement visibles qu'après déploiement.

[78] *Op. cit.* van der Weij et al. (2024).
[79] "Affaire Volkswagen." *Wikipédia, l'encyclopédie libre* (mai 2025). https://fr.wikipedia.org/wiki/Affaire_Volkswagen.
[80] *Op. cit.* Hendrycks. (2023).

Suite possible à cette note

1. Quelles sont les conséquences possibles et probables du développement d'une AGI dont le contrôle ne serait pas garanti ?
2. Quels niveaux de capacités ou de ressources rendent, en pratique, la correction ou la désactivation d'une AGI non réalisables ?

GPAI Policy Lab