

# Fonds d'alignement Demande de propositions

## Nos objectifs de recherche

Nous souhaitons financer des projets visant à atténuer les risques pour la sûreté et la sécurité découlant de systèmes d'intelligence artificielle (IA) désalignés. Les projets admissibles chercheront à relever l'un des défis suivants :

1. Comment peut-on empêcher les systèmes d'IA d'entreprendre ou de tenter d'entreprendre des actions qui constituent un risque pour notre sécurité collective?
2. Comment peut-on concevoir des systèmes d'IA qui n'essaient pas d'effectuer de telles actions en premier lieu?

Réaliser des avancées significatives dans ces domaines relève d'un effort interdisciplinaire qui nécessite une pluralité d'outils et de perspectives. Nous voulons que les plus brillants cerveaux dans de nombreux domaines contribuent à la recherche sur l'alignement. C'est pourquoi nous avons organisé cette demande de propositions sous la forme d'une série de questions propres à chaque discipline. Nous vous suggérons de cliquer directement sur les domaines qui vous intéressent plutôt que de lire le document dans son intégralité. Les sections sont classées approximativement de la plus théorique à la plus empirique.

Certaines des sous-sections ci-dessous fournissent davantage de détails que d'autres concernant les sous-problèmes, les recherches récentes et les recherches connexes. Il ne faut pas y voir une hiérarchisation des domaines que nous jugeons plus importants : cette disparité s'explique en grande partie par le fait que nous ou nos collaborateurs nous sommes concentrés jusqu'à présent sur certains domaines. Nous avons l'intention d'apporter des précisions semblables dans d'autres domaines, et nous nous efforcerons de le faire dans les prochaines versions de cette demande de propositions.

Nous sommes enthousiasmés par les projets qui traitent de ces questions, même s'ils ne sont pas axés sur un problème décrit ci-dessous. N'hésitez pas à consulter les listes et aperçus d'autres personnes – p. ex., [Google DeepMind](#), [Anthropic](#) et [Redwood Research](#) – pour trouver des idées. Si vous constatez des liens entre vos recherches et ces défis, nous vous encourageons à soumettre une proposition.

[Théorie de l'information et cryptographie](#)

[Théorie de la complexité](#)

[Théorie économique et théorie des jeux](#)

[Méthodes probabilistes](#)

[Théorie de l'apprentissage](#)

[Évaluation et garanties dans l'apprentissage par renforcement](#)

[Science cognitive](#)

[Interprétabilité](#)

[Conception et évaluation de tests de performances](#)

[Méthodes pour le post-entraînement et le déclenchement](#)

[Recherches empiriques sur la surveillance de l'IA et la méthode de l'équipe rouge](#)

# Domaines de recherche classés par discipline

## Théorie de l'information et cryptographie

La théorie de l'information fournit l'occasion de mettre au point des outils mathématiques permettant de comprendre les limites fondamentales du suivi et de la vérification des systèmes d'IA. Ces outils pourraient nous aider à déterminer ce qui peut et ne peut pas être détecté à propos du comportement de l'IA – si les modèles peuvent cacher de l'information dans leurs sorties, quand les portes dérobées peuvent être découvertes, et quel niveau de surveillance est théoriquement possible. En prouvant à la fois des résultats positifs (protocoles qui fonctionnent) et des résultats d'impossibilité (obstacles fondamentaux), la théorie de l'information peut aider à concentrer les efforts d'alignement sur des approches traitables. Les problèmes présentés dans cette section appliquent des outils comme l'entropie, la capacité des canaux et la difficulté cryptographique aux principaux défis de l'alignement : détecter la stéganographie dans les sorties des modèles, repérer les portes dérobées dans les systèmes entraînés et comprendre quand la surveillance peut échouer. À mesure que les systèmes d'IA gagnent en performance, ces limites théoriques deviennent cruciales pour concevoir des méthodes de surveillance qui ne peuvent être contournées.

*Approches relatives à la stéganographie et au raisonnement encodé fondées sur la théorie de l'information*

**Résumé du problème :** Il est théoriquement possible pour les systèmes d'IA d'encoder de l'information dans leurs sorties, tant de manière cryptographique que stéganographique. Peut-on théoriquement limiter cette possibilité et déterminer les approches susceptibles de la prévenir?

**Pourquoi cela est important :** Une mesure visant à empêcher les systèmes d'IA de causer un préjudice consiste à surveiller toute l'information et tous les raisonnements qu'ils produisent. Cela comprend à la fois la communication avec les êtres humains ou d'autres systèmes d'IA et l'utilisation d'un raisonnement de type « chaîne de résolution » pour améliorer leurs propres sorties. Le fait de cacher du contenu dans ces messages à l'aide de techniques de stéganographie ou de cryptographie compromettrait l'efficacité de notre surveillance.

**Exemples de sous-problèmes :**

1. Existe-t-il des limites théoriques pertinentes à ce qui peut être encodé de manière stéganographique ou cryptographique dans le langage naturel?

2. Quelles limites pourraient entraver la capacité des modèles à effectuer cet encodage dans la pratique? Les méthodes de stéganographie robuste sur le plan cryptographique nécessitent généralement des circuits beaucoup plus complexes que l'évaluation de l'espace latent d'un grand modèle de langage (GML). Il est donc difficile de savoir si un réseau *pourrait* apprendre de tels circuits et si la descente de gradient l'y inciterait.
3. Quelles mesures d'atténuation pourraient fonctionner? Est-il possible de mettre au point des limites théoriques appropriées pour les mesures d'atténuation existantes, comme la paraphrase ou la surveillance?
4. Si un signal faible de stéganographie est détecté, comment peut-on en tirer des preuves plus solides? Par exemple, il peut être possible d'obtenir d'un modèle de langage qu'il traduise occasionnellement une chaîne de résolution stéganographique en langage clair, mais avec une forte proportion de faux positifs : des affirmations selon lesquelles le modèle agit mal alors que ce n'est pas le cas. Existe-t-il des moyens fiables de filtrer les faux positifs afin de déterminer si l'on a affaire à de la stéganographie?

**Pour en apprendre davantage :** Parmi les premières études consacrées à ce problème, notons [Lanham et coll. \(2023\)](#), [Roger et Greenblatt \(2023\)](#) et [Turpin et coll. \(2023\)](#), dont nous recommandons la lecture. Les recherches antérieures dans ce domaine comprennent aussi [de Witt et coll. \(2023\)](#) et [Motwani et coll. \(2024\)](#). Bien que leurs objectifs soient différents, la détection ou la confirmation du recours à la stéganographie s'apparentent aux méthodes de tatouage numérique, comme [SynthID](#).

Parmi les recherches empiriques récentes, on peut citer [Skaf et coll. \(2025\)](#) et [Karpov et coll. \(2025\)](#), qui montrent que des modèles peuvent développer spontanément des chiffrements. Pour connaître les fondements théoriques de la stéganographie, se reporter à [Cachin \(2004\)](#) et à [Dai et Cai \(2019\)](#).

#### *Approches relatives aux portes dérobées fondées sur la théorie de l'information*

**Résumé du problème :** Prenons l'exemple d'un système d'IA modifié pour se comporter différemment en fonction d'une entrée « déclencheuse » particulière, connue sous le nom de « porte dérobée ». Il peut s'avérer impossible de détecter l'existence d'une telle porte dérobée avant que le système ne soit exposé à l'entrée « déclencheuse ». Existe-t-il des moyens de repérer ou d'atténuer les portes dérobées?

**Pourquoi cela est important :** Les systèmes d'IA peuvent agir comme on le souhaite pendant les phases d'entraînement et d'évaluation, mais de manière préjudiciable lorsqu'ils sont exposés à des conditions précises pendant le déploiement. Ces

conditions peuvent être délibérément définies pour ne se déclencher qu'une fois le système déployé. Il est essentiel d'évaluer si un modèle possède cette propriété afin d'empêcher le déploiement de systèmes préjudiciables.

#### Exemples de sous-problèmes :

1. Peut-on améliorer les notions formelles liées à la capacité de se défendre contre les portes dérobées?
2. Quand est-il possible de se protéger contre les portes dérobées?
3. Les portes dérobées peuvent-elles être supprimées moyennant un entraînement supplémentaire?
4. Quelles autres méthodes pourraient être utilisées pour contrer les portes dérobées?
5. Un défi théorique central doit être relevé dans le cadre des études portant sur la défense contre les portes dérobées dans les systèmes d'IA. Il s'agit de déterminer s'il existe une classe de représentation naturelle pouvant être défendue efficacement, mais ne pouvant pas être apprise de manière prouvable approximativement correcte (PAC). Il importe de cerner cette classe de représentation. Se reporter à [Christiano et coll. \(2024\)](#) pour plus de détails.
6. [Goldwasser et coll. \(2022\)](#) insèrent des portes dérobées indétectables en se fondant sur la difficulté de l'apprentissage par les erreurs. Cela nécessite toutefois un réseau doté d'une structure algébrique spéciale : un réseau utilisant le cosinus pour la périodicité. Existe-t-il une méthode plus générale et indétectable pour insérer une porte dérobée dont la structure est moins imposée?

**Pour en apprendre davantage :** [Goldwasser et coll. \(2022\)](#) fournit des résultats précurseurs et un aperçu général. Des avancées supplémentaires ont été réalisées par [Christiano et coll. \(2024\)](#), [Velegkas et coll. \(2024\)](#), [Goldwasser et coll. \(2025\)](#) et [Gluch et Goldwasser \(2025\)](#).

Pour connaître des études empiriques fondamentales, se reporter à [Gu et coll. \(2017\)](#), qui s'est penchée sur les attaques par insertion de déclencheurs, et à [Wang et coll. \(2019\)](#).

## Théorie de la complexité

Pour réaliser un alignement hautement fiable, il faudra fournir des garanties formelles et des observations empiriques. Tout comme le concept d'entropie a rendu la théorie de l'information beaucoup plus facile à traiter, nous soupçonnons l'existence de concepts intuitifs dans le domaine de la sécurité de l'IA qui, s'ils étaient formalisés mathématiquement, pourraient rendre soudainement traitables des problèmes jusqu'alors impossibles à traiter. Par exemple, de nombreuses approches visant à entraîner une IA sécuritaire reposent sur des êtres humains ou des systèmes d'IA plus faibles qui supervisent des systèmes plus performants. La théorie de la complexité aide à créer de telles approches et à les évaluer : dans quelles conditions présentent-elles des équilibres souhaitables, quand peut-on prouver qu'elles sont impossibles et comment peut-on en limiter les erreurs ?

*Approches fondées sur la théorie de la complexité visant à améliorer les protocoles de débat*

**Résumé du problème :** La sécurité de l'IA par le débat ([Irving et coll., 2018](#)) constitue une méthode potentielle pour entraîner sécuritairement une IA avancée afin qu'elle comprenne les objectifs complexes des êtres humains. Les agents sont entraînés à l'aide d'un jeu de débat à somme nulle, au cours duquel un être humain juge lequel des agents a fourni l'information la plus véridique et la plus utile. Il existe d'autres protocoles de débat, comme le contre-interrogatoire ([Barnes et Christiano, 2020](#)), le débat doublement efficace ([Brown-Cohen et coll., 2023](#)) et le débat entre démonstrateur et estimateur ([Brown-Cohen et coll., 2025](#)). Tous ces protocoles sont insuffisants à au moins un égard ([Buhl et coll., 2025](#)) ; peut-on les améliorer ?

**Pourquoi cela est important :** La sécurité de l'IA par le débat cherche à résoudre le problème de la surveillance évolutive. Cette approche consiste à récompenser correctement les comportements souhaités des systèmes d'IA, même lorsque ces comportements dépassent la capacité des êtres humains à les juger efficacement. Une méthode de surveillance évolutive robuste est nécessaire pour entraîner en toute sécurité des modèles dotés de capacités très avancées, et le débat fait partie des approches les plus prometteuses.

**Exemples de sous-problèmes :**

1. Dans le débat entre démonstrateur et estimateur ([Brown-Cohen et coll., 2025](#)), la complexité du circuit de l'estimateur dépend de divers facteurs liés à la profondeur, au degré de ramification, à la stabilité et à la limite d'erreur souhaitée. Peut-on supprimer l'un de ces facteurs ou montrer que la limite est

serrée? Cela se traduit par de nombreux sous-problèmes possibles, étant donné que les différents facteurs apparaissent pour différentes raisons.

2. Dans la première version de la sécurité de l'IA par le débat ([Irving et coll., 2018](#)), des joueurs en concurrence pouvaient fournir des contre-arguments aux déclarations de leurs adversaires. Cette notion est écartée pour ce qui est du débat entre démonstrateur et estimateur. Est-il possible d'élaborer un protocole amélioré qui réintroduise les contre-arguments – et, par extension, le contre-interrogatoire – tout en conservant les avantages procurés par le débat entre démonstrateur et estimateur, comme le fait d'éviter les [arguments obscurs](#)? Les techniques ou les résultats issus des preuves interactives multidémonstrateurs peuvent-ils être appliqués à ce paramètre?
3. La preuve de l'exhaustivité du débat entre démonstrateur et estimateur ([théorème 6.1](#)) exige que le démonstrateur soit en mesure d'estimer les probabilités avec précision afin de savoir à quel endroit il doit procéder à la récursivité. Peut-on réduire cette exigence de précision?
4. La preuve d'existence pour l'estimateur dans le débat entre démonstrateur et estimateur ([théorème 6.1](#)) repose sur un algorithme à temps exponentiel. Nous estimons que cela est probablement nécessaire; il faut prouver l'affirmation ou fournir un contre-exemple. De plus, il faut trouver des hypothèses naturelles pour les preuves qui pourraient admettre une version en temps polynomial de l'algorithme. La stabilité ( $\epsilon, \rho$ ) est-elle la bonne notion et existe-t-il une meilleure définition de la stabilité? La définition utilisée dans les preuves pour le débat entre démonstrateur et estimateur est autorégressive en raison de la manière dont nous obtenons la convexité pour la descente de gradient en ligne, ce qui se traduit ensuite par un facteur de  $2^q$  dans la complexité du démonstrateur.
5. Dans la pratique, la plupart des calculs impliquent au moins un point instable. Par exemple, on pourrait approfondir le problème avant de présenter plusieurs éléments de preuve, et une mauvaise analyse pourrait compromettre l'ensemble du calcul. Dans les théorèmes relatifs au débat entre démonstrateur et estimateur, on modélise ces parties instables du calcul comme étant connues de tout le monde, y compris de l'estimateur, mais cela n'est pas réaliste dans le cas des débats en langage naturel. Cependant, on prévoit que le coût d'une erreur sera proportionnel à son importance stratégique : les erreurs accidentelles sont préférables aux erreurs intentionnelles. Il faut formaliser l'intuition selon laquelle éviter les erreurs intentionnelles nécessite que les propriétés de stabilité du calcul soient plus faibles.

6. Le débat entre démonstrateur et estimateur s'appuie sur la notion d'« éléments de preuve multiples et indépendants » pour stabiliser le débat. Dans divers contextes théoriques, il existe des théorèmes de seuil montrant comment transformer des calculs instables en calculs stables, p. ex., la correction d'erreurs dans les calculs. Peut-on élaborer des théorèmes connexes pour le contexte du langage naturel?
7. Le débat entre démonstrateur et estimateur repose sur le fait que l'estimateur dispose d'un certain facteur plus calculable que celui du démonstrateur. Peut-on mettre au point des protocoles qui sont honnêtes ou exacts de manière prouvable dans un état d'équilibre, malgré une limitation supplémentaire des ressources de calcul disponibles pour les joueurs?
8. On parle de bidouillage exploratoire lorsque, dans un débat, un joueur refuse de donner sa meilleure réponse afin d'empêcher que celle-ci soit échantillonnée puis renforcée par le processus d'entraînement. Il faut trouver un autre protocole qui prévient cela et prouver son exactitude. Par exemple, un protocole pourrait autoriser les débatteurs à proposer des extensions pour corriger les solutions, sans toutefois les optimiser.
9. Les erreurs humaines systématiques constituent un problème majeur quant à l'exactitude des sorties des débats. Est-il possible de produire des protocoles de débat entre démonstrateur et estimateur dans lesquels ce dernier est en mesure d'empêcher les requêtes à l'oracle qui exploitent les biais humains systématiques ([Irving, 2025](#))?
10. Existe-t-il des analogies avec les résultats du débat entre démonstrateur et estimateur dans des versions plus intrinsèquement bayésiennes de la surveillance évolutive, comme [Bengio et coll. \(2025\)](#)? Même l'inférence bayésienne élémentaire sur des graphes de taille polynomiale requiert généralement un temps exponentiel. Par conséquent, il est probable que les résultats formels nécessiteront des critères de réussite moins stricts, par exemple en exigeant des agents qu'ils obtiennent des résultats indiscernables de ceux du vrai postérieur.

**Pour en apprendre davantage :** Nous recommandons la lecture de [Buhl et coll. \(2025\)](#), qui explique comment l'approche du débat pourrait convaincre qu'un système d'IA est sûr, et [Brown-Cohen et coll. \(2025\)](#), qui présente un protocole récent relatif au débat. Les recherches antérieures dans ce domaine comprennent [Irving et coll. \(2018\)](#), [Barnes et Christiano \(2020\)](#), [Brown-Cohen et coll. \(2023\)](#), ainsi que [Brown-Cohen et Irving](#)

(2024). [Leike et coll. \(2018\)](#) lancent l'idée générale d'une modélisation récursive des récompenses, au-delà du sujet du débat.

### *Indépendance des preuves*

**Résumé du problème :** Les systèmes d'IA auxquels nous faisons appel pour résoudre des problèmes complexes doivent proposer et défendre différentes affirmations concernant la validité ou la sécurité de leurs solutions. Pour vérifier ces affirmations, nous aurons besoin de preuves, souvent fournies par les IA elles-mêmes. Ces preuves sont généralement imprécises ou probabilistes, ce qui nous oblige à combiner plusieurs sources pour parvenir à une conclusion fiable. Cependant, il est difficile d'évaluer si plusieurs éléments de preuve sont véritablement indépendants (offrant une redondance solide) ou s'ils sont en réalité corrélés ou proviennent de la même source sous-jacente. Cela crée un mode de défaillance critique : on pourrait avoir une confiance excessive dans une affirmation parce qu'une IA fournit de nombreux éléments de preuve qui semblent indépendants, mais qui sont de fait corrélés. Bien qu'il s'agisse d'un problème général dans le raisonnement, il est particulièrement crucial pour la validité des méthodes de surveillance évolutive.

**Pourquoi cela est important :** Il est nécessaire de déterminer l'indépendance afin d'évaluer correctement notre confiance envers les solutions ou les arguments proposés par l'IA. Par exemple, dans l'étude théorique des méthodes de surveillance évolutive, nous devons parfois supposer l'existence d'éléments de preuve non corrélés pour attester de la validité de la méthode, p. ex., en supposant un [soutien s dans le débat entre démonstrateur et estimateur](#). Nous nous attendons également à ce que cet enjeu se pose de manière empirique, et il pourrait être particulièrement préjudiciable à la surveillance humaine de l'IA, compte tenu des nombreux biais systématiques dont font preuve les juges humains. Ces biais sont fortement corrélés dans le temps et entre les différents juges, plutôt que d'être des erreurs indépendamment distribuées.

### **Exemples de sous-problèmes :**

1. Comment peut-on déterminer dans la pratique si différents éléments de preuve sont indépendants? Cela dépend probablement beaucoup du contexte. C'est pourquoi nous recommandons de se concentrer sur les domaines où l'IA est susceptible d'être la plus utilisée et où la précision épistémique est requise, comme le génie, les mathématiques ou la recherche scientifique. La bonne notion peut également dépendre du caractère empirique (consultation de vastes bases de données expérimentales) ou conceptuel (débat abstrait portant sur la philosophie ou les mathématiques) des preuves.

2. Comment devrait-on modéliser l'indépendance dans l'étude théorique des méthodes de surveillance évolutive? En particulier, quelle formulation de l'indépendance est nécessaire pour prouver la validité d'une méthode? Par exemple, existe-t-il d'autres notions d'indépendance que celle actuellement employée dans [Brown-Cohen et coll. \(2025\)](#) qui pourraient encore garantir la stabilité ( $\epsilon$ ,  $\rho$ )?
3. Pour quels sujets ces notions d'indépendance devraient-elles s'appliquer, comme le soutien  $s$  dans [Brown-Cohen et coll. \(2025\)](#)? Comment peut-on vérifier de manière empirique si c'est le cas pour un sujet d'intérêt?
4. Peut-on utiliser de manière récursive des méthodes de surveillance évolutive pour résoudre les questions relatives à l'indépendance? Il existe peut-être des méthodes spéciales particulièrement efficaces pour résoudre ce type de questions, même si elles le sont moins pour d'autres.
5. Comment le problème de l'indépendance des preuves évolue-t-il dans des versions plus intrinsèquement bayésiennes de la surveillance évolutive, comme celles proposées par [Bengio et coll. \(2025\)](#)? Peut-on s'appuyer sur des notions plus floues d'indépendance ou réduire autrement le fardeau?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Brown-Cohen et coll. \(2025\)](#) et de [Bengio et coll. \(2025\)](#), qui étend la surveillance évolutive au cadre bayésien.

#### *Jeux entre démonstrateur et vérificateur*

**Résumé du problème :** Une formulation du problème de la surveillance évolutive consiste à savoir si un *vérificateur* digne de confiance aux capacités de calcul limitées peut apprendre à résoudre des tâches en interagissant avec un *démonstrateur* plus puissant, mais non digne de confiance ([Brown-Cohen et coll., 2025](#)).

**Pourquoi cela est important :** À mesure que les capacités de l'IA s'améliorent, nous devrons pouvoir surveiller les sorties des modèles que nous ne pourrons pas évaluer directement. Dans ce contexte, le recours à des modèles d'IA pour surveiller d'autres modèles d'IA est un outil essentiel pour fournir des signaux d'entraînement vers le développement de traits souhaitables, comme l'honnêteté.

#### **Exemples de sous-problèmes :**

1. Existe-t-il de bons protocoles pour entraîner un démonstrateur?

2. Quelles garanties théoriques peut-on obtenir quant à l'efficacité des paramètres des jeux entre démonstrateur et vérificateur? Pour quels problèmes peut-on obtenir des garanties prouvables? À quel point ces problèmes correspondent-ils à des scénarios de déploiement réels?
3. Peut-on trouver des problèmes pour lesquels un système démonstrateur-vérificateur ne peut garantir la sécurité de manière prouvable, et qui correspondent aux problèmes que nous souhaitons résoudre dans des scénarios de déploiement?
4. Le schéma de base des jeux entre démonstrateur et vérificateur peut-il être amélioré afin de rendre la procédure plus puissante en théorie ou plus réalisable dans la pratique?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Hammond et Adam-Day \(2025\)](#). Les recherches antérieures dans ce domaine comprennent [Kirschner et coll. \(2024\)](#) et [Amit et coll. \(2024\)](#).

#### *Petits circuits superposés à des fonctions heuristiques*

**Résumé du problème :** Les futures techniques d'interprétabilité pourraient nous permettre d'acquérir une compréhension générale des circuits d'un modèle, sans toutefois comprendre entièrement les sous-processus. On pourrait modéliser cette situation comme un petit circuit faisant appel à des fonctions heuristiques. Un tel cadre a constitué un modèle théorique efficace dans plusieurs domaines de la théorie de la complexité, notamment la confidentialité différentielle et l'équité statistique.

**Pourquoi cela est important :** Un certain nombre d'outils de sécurité de l'IA actuels pourraient être modélisés à l'aide de petits circuits et de fonctions heuristiques. Si nous parvenons à mieux comprendre ces outils, nous pourrions être en mesure de déterminer quand ils fonctionnent ou non. Par exemple, on pourrait comprendre quand faire confiance à nos outils d'interprétabilité ou quelles interventions pourraient améliorer l'interprétabilité.

#### **Exemples de sous-problèmes :**

1. Quels paramètres de sécurité de l'IA peuvent être modélisés sous forme de petits circuits de niveau supérieur dotés de fonctions de niveau inférieur?
2. Étant donné que la plupart des techniques de chiffrement nécessitent des circuits complexes, les résultats existants concernant la difficulté de détection des portes dérobées changerait-ils si l'on se fondait sur ce modèle de circuits simples? Qu'est-ce que cela nous enseigne au sujet de la capacité de l'interprétabilité à déclencher des contextes néfastes?

**Pour en apprendre davantage :** Pour découvrir un exemple concret où les petits circuits se sont déjà révélés utiles, se reporter à [Brown-Cohen et coll. \(2025\)](#), qui présente un cas d'utilisation relatif à l'alignement. Pour d'autres cas d'utilisation, se reporter à [Dwork et coll. \(2022a, 2022b\)](#). Une application potentielle est décrite dans [Irving et coll. \(2025\)](#).

#### *Prouvable correct vs prouvable honnête*

**Résumé du problème :** Les théorèmes relatifs au débat entre démonstrateur et estimateur ([Brown-Cohen et coll., 2025](#)) montrent qu'à l'état d'équilibre, la plupart du temps, les affirmations du démonstrateur sont correctes ([théorème 6.1](#)). Intuitivement, l'honnêteté est nécessaire, mais pas suffisante pour garantir l'exactitude. La formalisation de cette notion pourrait permettre d'affaiblir la stabilité et les besoins en ressources pour le débat entre démonstrateur et estimateur. Il existe différentes propriétés de l'honnêteté que l'on pourrait examiner.

**Pourquoi cela est important :** Dans la pratique, on ne sait pas à l'avance si un problème donné peut être résolu correctement au moyen d'un débat. Pour les problèmes auxquels il n'existe pas de réponse parfaite, les théorèmes actuels n'offrent aucune garantie : le débat pourrait déboucher sur n'importe quel résultat. Si l'on pouvait définir l'« honnêteté » comme une propriété moins importante que l'exactitude, on pourrait garantir que le débat aboutisse à la meilleure estimation possible du modèle, même lorsque la certitude est impossible, tout en conservant des garanties de sécurité.

#### **Exemples de sous-problèmes :**

1. Quelles définitions formelles de l'honnêteté pourraient être utiles? Notre hypothèse initiale, en employant la notation de [Brown-Cohen et coll. \(2025\)](#), serait « Tous les petits circuits  $C(x)$  sont au plus de l'ordre de  $O(\epsilon)$  meilleurs que le hasard pour distinguer les exécutions correctes de  $M(x)$  des exécutions incorrectes ».
2. Comment celles-ci nous permettent-elles de réduire les exigences en matière de stabilité et de ressources? Y a-t-il une définition de l'honnêteté qui favorise un débat plus efficace entre démonstrateur et estimateur?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Brown-Cohen et coll. \(2025\)](#). Les recherches antérieures dans ce domaine comprennent [Amit et coll. \(2024\)](#) et [Neyman \(2024\)](#).

#### *Approches relatives à la recherche inexploitable fondées sur la théorie de la complexité*

**Résumé du problème :** Au fil du temps, si un signal de récompense dans le cadre d'un entraînement admet plusieurs solutions, ces choix libres risquent d'être exploités (le « problème de la recherche exploitable »); [Pfau et Irving, 2025](#)). Les méthodes de

recherche inexploitables devraient contraindre un modèle à répartir aléatoirement ses performances entre tous les objectifs secondaires antagonistes auxquels il a accès, dans un sens qui serait formalisé par la méthode.

**Pourquoi cela est important :** Nous aimerais pouvoir confier à l'IA des tâches insuffisamment spécifiées, comme le conseil en recherche, pour lesquelles il n'existe pas de solution correcte vérifiable de manière unique. Dans ce contexte, nous pourrions être vulnérables à un système d'IA qui sélectionnerait de façon antagoniste des solutions plausibles, mais néanmoins préjudiciables, comme l'ouverture d'une surface pour de futures exploitations ou la conservation d'information stéganographique en vue d'un comportement répréhensible ultérieur.

#### **Exemples de sous-problèmes :**

1. Comment peut-on restreindre l'espace des solutions selon l'orientation pertinente? Par exemple, le hachage, qui exige qu'une solution en langage naturel ait une certaine valeur de hachage, offre une méthode évidente pour réduire l'espace des paramètres libres. Cependant, une application naïve du hachage du texte d'une solution donnée ne ferait qu'obliger le modèle à itérer sur la formulation jusqu'à ce que la solution souhaitée soit trouvée. Il importe d'isoler les éléments clés de l'espace à restreindre.
2. Sinon, peut-on prouver qu'un système n'utilise pas les paramètres libres, c.-à-d. qu'il choisit une option arbitraire parmi des options tout aussi valables les unes que les autres?
3. S'il est possible de limiter les paramètres libres afin d'empêcher toute recherche exploitante, comment peut-on créer des restrictions qui résistent aux comportements volontairement trompeurs (*sandbagging*)? Par cette manœuvre, un modèle sous-performe de façon intentionnelle afin de nous amener à restreindre insuffisamment l'espace.

**Pour en apprendre davantage :** Nous recommandons la lecture de [Pfau et Irving \(2025\)](#).

#### *Approches relatives aux portes dérobées fondées sur la théorie de la complexité*

**Résumé du problème :** Prenons l'exemple d'un système d'IA modifié pour se comporter différemment en fonction d'une entrée « déclencheuse » particulière, connue sous le nom de « porte dérobée ». Il peut s'avérer impossible de détecter l'existence d'une telle porte dérobée avant que le système ne soit exposé à l'entrée « déclencheuse ». Existe-t-il des moyens de repérer ou d'atténuer les portes dérobées?

**Pourquoi cela est important :** Les systèmes d'IA peuvent agir comme on le souhaite pendant les phases d'entraînement et d'évaluation, mais de manière préjudiciable

lorsqu'ils sont exposés à des conditions précises pendant le déploiement. Ces conditions peuvent être délibérément définies pour ne se déclencher qu'une fois le système déployé. Il est essentiel d'évaluer si un modèle possède cette propriété afin d'empêcher le déploiement de systèmes préjudiciables.

### **Exemples de sous-problèmes :**

#### *Sous-problèmes établis :*

1. Peut-on améliorer les notions formelles liées à la capacité de se défendre contre les portes dérobées, p. ex., [Goldwasser et coll. \(2025\)](#)?
2. Quand est-il possible de se protéger contre les portes dérobées?
3. Quelles autres méthodes pourraient être utilisées pour contrer les portes dérobées?
4. Un défi théorique central à relever dans le cadre des recherches portant sur la défense contre les portes dérobées dans les systèmes d'IA consiste à déterminer s'il existe une classe de représentation naturelle pouvant être défendue efficacement, mais ne pouvant pas être apprise de manière PAC. Il importe de cerner cette classe de représentation. Se reporter à [Christiano et coll. \(2024\)](#) pour plus de détails.
5. Les portes dérobées peuvent-elles constituer une mesure de sécurité, par exemple en empêchant le modèle de prendre en compte certains plans préjudiciables?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Christiano et coll. \(2024\)](#). Les recherches antérieures dans ce domaine comprennent [Goldwasser et coll. \(2022, 2025\)](#) et [Gluch et Goldwasser \(2025\)](#).

#### *Arguments heuristiques*

**Résumé du problème :** Peut-on formaliser le concept d'« argument heuristique » – argument informel portant sur le comportement d'un système mathématique – afin de le rendre vérifiable par une machine? Par exemple, si l'on suppose que certains faits sont non corrélés ou indépendants, peut-on tirer des conclusions utiles pour entraîner des systèmes d'IA sécuritaires?

**Pourquoi cela est important :** Bon nombre des arguments que l'on pourrait avancer à propos des systèmes d'IA, par exemple concernant leurs structures internes ou la probabilité de défaillances rares, seront des heuristiques informelles ou probabilistes, plutôt que des preuves irréfutables. Nous aimerions pouvoir recourir à des arguments heuristiques vérifiables par une machine, mais pas nécessairement compréhensibles

par un être humain, afin de formuler des affirmations prudentes, mais non certaines, au sujet des systèmes d'IA.

### Exemples de sous-problèmes :

1. Peut-on prouver la conjecture de non-coïncidence computationnelle ([Neyman, 2025](#)) – une formalisation de l'idée selon laquelle « si une coïncidence apparemment scandaleuse se produit en mathématiques, c'est qu'il y a une raison à cela »?
2. The ARC problems on
3. Quelles sont les méthodes d'estimation de probabilités rares, qu'elles soient fondées ou non sur l'échantillonnage? L'Alignment Research Center (Centre de recherche sur l'alignement) soupçonne que les méthodes sans échantillonnage pourraient connaître une expansion importante. On peut parler aussi bien de méthodes pratiques que d'algorithmes impraticables, mais intéressants sur le plan théorique (p. ex., même un algorithme galactique serait génial).
4. L'Alignment Research Center avait déjà présenté deux problèmes récompensés d'un prix concernant le remplissage de matrices :  
<https://www.alignment.org/blog/prize-for-matrix-completion-problems>. Je crois que ces questions sont toujours pertinentes, et nous devrions les inclure si c'est le cas.
5. À un moment donné, l'Alignment Research Center disposait d'une liste très précise des propriétés que devaient posséder certaines fonctions ou certains ensembles de neurones afin de répondre à ses objectifs de formalisation des arguments heuristiques. Nous devrions déterminer s'ils ont jamais rédigé cette liste et, si tel est le cas, ajouter un lien vers celle-ci. Si ce n'est pas le cas, nous devons leur demander de la consigner par écrit. C'est un mauvais mode de défaillance si les partisans de la théorie de l'alignement ne montrent pas leur travail, y compris les axes de recherche qui sont dans une impasse.

**Pour en apprendre davantage :** Nous recommandons la lecture de [Hilton \(2024\)](#) pour obtenir une analyse plus approfondie du problème et un résumé de certains résultats, ainsi que de [Christiano et coll. \(2022\)](#) et de [Neyman \(2025\)](#) pour connaître les résultats en la matière.

*Des projets audacieux qui définissent de nouveaux paradigmes pour comprendre la difficulté de l'alignement*

**Résumé du problème :** Existe-t-il d'autres approches pour créer et déployer des systèmes d'IA avancés qui atténuent de manière crédible les risques liés à des

systèmes désalignés? Dans quelle mesure ces approches seraient-elles difficiles à mettre en œuvre malgré des efforts concertés?

**Pourquoi cela est important :** Nous ne savons pas vraiment si les approches actuelles fonctionneront et si elles pourront être adaptées à des systèmes d'IA encore plus puissants.

**Pour en apprendre davantage :** Nous recommandons la lecture de [Hilton \(2024\)](#), d'[Appel et Kosoy \(2025\)](#), ainsi que de [Nayebi \(2025\)](#).

## Théorie économique et théorie des jeux

Nous nous réjouissons d'utiliser les cadres et les outils de l'économie théorique pour mieux décrire les mesures incitatives et les résultats d'équilibre pertinents pour l'alignement de l'IA. En considérant les agents d'IA dans un contexte stratégique, nous espérons nous assurer que les approches d'alignement sont suffisamment robustes pour résister aux améliorations des capacités et que ces cadres peuvent inspirer de nouvelles idées de techniques d'alignement. Nous prévoyons que la proposition de nouvelles abstractions ou d'abstractions différentes permettra d'obtenir des enseignements utiles : comment différentes formes de mesures incitatives désalignées influent sur les résultats ou à quel point les propositions sont solides face à différents concepts d'équilibre.

[Hammond et coll. \(2025\)](#) fournissent une taxonomie des risques liés à l'IA avancée dans des environnements multiagents, y compris une liste plus complète des problèmes et un contexte plus large pour certains des problèmes précis que nous soulevons ci-dessous.

### *Conception de l'information*

**Résumé du problème :** Les modèles d'avant-garde actuels présentent déjà des comportements [flagorneurs](#) et [trompeurs](#), qui pourraient être considérés comme des preuves d'un comportement stratégique. À mesure que les modèles d'IA deviennent plus stratégiques, comment devraient-ils choisir l'information à divulguer? Nous espérons tirer des enseignements utiles de domaines liés à la conception de l'information, comme le verbiage ([Crawford et Sobel, 1982](#)), la divulgation d'information vérifiable ([Milgrom, 1981](#)), la persuasion ([Kamenica et Gentzkow, 2011](#)) ou la délégation ([Holmstrom, 1984](#)). Quelles sont les stratégies de divulgation optimales du point de vue de l'IA, dans quelle mesure les équilibres sont-ils variés et stables, et peut-on concevoir des mécanismes pour améliorer la divulgation d'information?

**Pourquoi cela est important :** La divulgation d'information revêt une importance pour plusieurs aspects de l'alignement. Certaines propositions en matière d'alignement se concentrent sur l'élaboration de protocoles visant à encourager l'honnêteté (p. ex., [Irving et coll., 2018](#)). Si l'on a confiance qu'une IA est honnête ou incapable de nous tromper, on peut compter sur elle pour superviser ou concevoir d'autres modèles (p. ex., [Buhl et coll., 2025](#)). De plus, quand peut-on employer une IA désalignée, même si elle tente de nous manipuler? (se reporter à [Bhatt et coll., 2025](#) pour obtenir une introduction à la configuration de base de l'évaluation des contrôles).

## Exemples de sous-problèmes :

1. Comment la divulgation optimale d'information (étant donné qu'une ou plusieurs IA peuvent avoir un objectif différent) interagit-elle avec les protocoles d'alignement par le débat ([Irving et coll., 2018](#))? Selon quelles hypothèses ces protocoles échouent-ils?
2. Contrairement à ce qui est généralement admis dans la littérature économique, la fonction utilitaire de l'émetteur (IA) peut être modifiée en fonction de ses actions précédentes (p. ex., par apprentissage par renforcement). Si une IA est consciente qu'elle est en cours d'entraînement (« conscience situationnelle »), elle peut vouloir éviter d'être modifiée. Comment la conscience situationnelle pourrait-elle modifier la dynamique du jeu du débat? Par exemple, [Chen et coll. \(2024\)](#) utilisent le rappel imparfait pour modéliser la conscience situationnelle imparfaite.
3. Si l'on sait qu'un modèle d'IA est désaligné, comment peut-on l'utiliser en toute sécurité? Cela pourrait être étudié en élargissant les cadres de manipulation dans les ouvrages sur la persuasion ([Kamenica et Gentzkow, 2011](#)) ou la conception de mécanismes dans lesquels l'espace contractuel représente les environnements d'entraînement possibles.
4. [Kamenica et Gentzkow \(2011\)](#) modélisent la transmission d'information lorsque l'émetteur tente de manipuler le récepteur. En supposant que nous (le récepteur) sachions qu'une IA désalignée tente de nous tromper, peut-on concevoir des mécanismes permettant d'améliorer la quantité d'information qu'il est optimal d'émettre (p. ex., en limitant la forme que doit prendre le message)? À quel point ce modèle est-il sensible à ses hypothèses (dimensions multiples, rationalité limitée, erreurs)?
5. Lorsque le fait de signaler la probabilité d'un événement modifie cette probabilité, que veut-on encourager? Les recherches antérieures dans ce domaine comprennent [Armstrong et O'Rorke \(2017\)](#), [Perdomo et coll. \(2020\)](#), ainsi que [Hudson \(2024\)](#).

**Pour en apprendre davantage :** Pour en savoir plus sur l'alignement par le débat, nous recommandons la lecture de [Buhl et coll. \(2025\)](#) et de [Brown-Cohen et coll. \(2025\)](#). Les recherches antérieures relatives au débat incluent [Irving et coll. \(2018\)](#), [Barnes \(2020\)](#), [Brown-Cohen et coll. \(2023\)](#), ainsi que [Brown-Cohen et Irving \(2024\)](#). Pour sa part, [Hammond et coll. \(2025\)](#) traite d'un large éventail de risques découlant de scénarios multiagents; se reporter en particulier à la section 3.1 sur les asymétries d'information.

### *Conception de mécanismes (robustes)*

**Résumé du problème :** À mesure que les modèles d'IA se perfectionnent, ils peuvent acquérir des capacités que nous ne prévoyons pas. Si nous modélisons l'alignement comme un jeu entre un principal humain et un agent d'IA, à quel point les mécanismes de sécurité proposés peuvent-ils résister à des actions inconnues dans l'espace d'action de l'IA?

**Pourquoi cela est important :** Nous voulons des méthodes d'alignement qui puissent s'adapter à l'augmentation des capacités futures, même si nous ne pouvons pas anticiper exactement celles-ci.

### **Exemples de sous-problèmes :**

1. [Carroll \(2015\)](#) démontre l'optimalité des contrats linéaires lorsque le principal ne connaît pas l'espace d'action de l'agent. Ce résultat repose sur l'hypothèse qu'aucune action (comme un sabotage d'entreprise) susceptible de nuire gravement au principal n'est entreprise. Nous ne voulons pas exclure de telles actions lorsque nous examinons les actions qu'un agent d'IA pourrait prendre. Peut-on trouver le mécanisme optimal pour le principal en tenant compte du compromis entre la réduction de la probabilité de préjudice et la réduction de l'utilité d'un modèle?
2. Comment devons-nous envisager l'espace d'action contractable ou non contractable lorsque nous (le principal) interagissons avec une IA (l'agent) ou lui déléguons des décisions (qui peuvent être désalignées)?

**Pour en apprendre davantage :** [Hadfield-Menell et Hadfield \(2018\)](#) ont défini certains cadres de l'alignement de l'IA comme étant un problème de contrat incomplet. Nous recommandons la lecture de [Bhatt et coll. \(2025\)](#), qui fournit une introduction à la configuration de base de l'évaluation des contrôles. Les recherches antérieures menées sur le contrôle comprennent [Mallen et coll. \(2025\)](#).

### *Rationalité limitée*

**Résumé du problème :** Nous voulons prédire comment les systèmes d'IA puissants se comporteront dans divers scénarios. Selon nos meilleures estimations, il s'agirait d'une combinaison entre l'extrapolation de comportements existants (à partir de modèles actuels aux capacités plus limitées), le calcul de stratégies optimales dans des équilibres de Nash (agents sans limite de calcul) et le calcul de stratégies optimales pour des agents avec limite de calcul (théorie de la complexité et théorie algorithmique

des jeux). Existe-t-il d'autres approches, inspirées de l'économie comportementale ou d'autres domaines, qui suggèrent un comportement significativement différent de la part des IA puissantes?

**Pourquoi cela est important :** Nous avons actuellement beaucoup d'incertitudes quant au fonctionnement des réseaux de neurones et de l'apprentissage profond, et donc quant à leur évolutivité. La plupart des études sur l'interprétabilité sont empiriques et s'appuient donc sur des changements de comportement passés pour extrapoler des modèles futurs. Afin de comprendre si les techniques de contrôle et d'alignement proposées fonctionneront, nous devons mieux saisir comment la rationalité limitée influera sur les actions qu'une IA pourrait entreprendre.

#### **Exemples de sous-problèmes :**

1. De nombreuses recherches sur l'alignement (p. ex., [Brown-Cohen et coll., 2023](#)) s'appuient sur la théorie de la complexité pour modéliser la rationalité limitée en termes de contraintes de calcul. D'autres formalisations de la rationalité limitée dans le jeu de débat suggèrent-elles des équilibres différents (par rapport à des agents limités ou illimités en termes de calcul)?
2. Il peut être difficile de juger de la sécurité des modèles d'IA qui sont considérablement plus performants ou « intelligents » que les êtres humains. La surveillance évolutive (également abordée ailleurs dans ce document) propose de recourir à une série de modèles d'IA afin d'aligner et de surveiller des modèles de plus en plus performants. Comment devrait-on modéliser « plus performant » ou « plus intelligent » dans ce cadre? Y a-t-il des conceptualisations qui, si elles s'avéraient, rendraient plus difficile, voire impossible, la réussite de ce cadre?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Buhl et coll. \(2025\)](#) et de [Brown-Cohen et coll. \(2025\)](#). Les recherches antérieures relatives au débat comprennent [Irving et coll. \(2018\)](#), [Barnes \(2020\)](#), [Brown-Cohen et coll. \(2023\)](#), ainsi que [Brown-Cohen et Irving \(2024\)](#). Il peut être particulièrement utile d'examiner les cas où le principal et l'agent sont tous deux limités, mais où le principal l'est considérablement plus que l'agent. Par exemple, [Goldreich \(2017\)](#) présente un système de preuve dans lequel le démonstrateur est moins soumis à des contraintes que le vérificateur.

#### ***Théorie des jeux à code source libre, connivence et engagement***

**Résumé du problème :** L'IA peut être plus apte à mener des actions concertées ou à s'engager dans des actions que les êtres humains. L'IA peut être en mesure de

partager une partie de son code ou de sa chaîne de résolution ([Wei et coll., 2022](#) et [Korbak et coll., 2025](#)), ce qui augmente sa capacité à transmettre des signaux crédibles d'engagement. En outre, la manière dont les IA agissent de connivence (p. ex., les actions qu'elles coordonnent en l'absence de communication) peut différer de la façon dont nous nous attendons à ce que les êtres humains agissent de connivence, ce qui rend cette dernière plus difficile à détecter.

**Pourquoi cela est important :** Certaines propositions d'alignement s'appuient sur des modèles d'IA indépendants dans des contextes antagonistes, par exemple en faisant surveiller les IA les unes par les autres ([Goel et coll., 2025](#)).

#### Exemples de sous-problèmes :

1. Des problèmes concrets non résolus relatifs à la théorie des jeux à code source libre sont présentés dans [Critch et coll. \(2022\)](#). Cet article examine les agents dont le code source est (partiellement) observable par les autres joueurs. Cela pourrait déterminer dans quelle mesure les agents doivent être interprétables ou observables les uns pour les autres (ou pour un être humain) afin d'encourager la coopération ou de décourager la connivence.
2. Comment peut-on concevoir des mécanismes pour empêcher la connivence entre les agents d'IA et un système de surveillance? (Se reporter à Bhatt et coll., 2025 pour avoir un exemple d'une configuration d'évaluation des contrôles.)
3. Quels concepts de solution faut-il adopter pour modéliser les interactions entre les IA? Par exemple, le perfectionnement des sous-jeux peut s'avérer trop restrictif lorsque les IA sont capables de signaler leur engagement de manière crédible. (Se reporter à [Tennenholtz, 2004](#) pour obtenir un exemple lié à la théorie des jeux à code source libre.)
4. Lorsque des modèles d'IA indépendants sont utilisés dans des propositions d'alignement ([Hammond et coll., 2025](#), page 43), quelle définition précise ou formalisation de l'indépendance est nécessaire pour que ces protocoles fonctionnent comme prévu?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Critch et coll. \(2022\)](#) et d'[Oesterheld et coll. \(2023\)](#). Les recherches antérieures portant sur la théorie des jeux à code source libre comprennent [Sauerberg et Oesterheld \(2025\)](#). [Hammond et coll. \(2025\)](#) traite de la connivence entre les IA avancées dans les sections 2.3 et 4.1 (on y retrouve une discussion sur la stéganographie ou les messages chiffrés). [Bhatt et](#)

coll. (2025) présente la configuration de base pour l'évaluation des contrôles, en s'appuyant sur les recherches antérieures de Mallen et coll. (2025).

## Méthodes probabilistes

### *Estimation de faible probabilité*

**Résumé du problème :** [Wu et Hilton \(2025\)](#) Wu et Hilton (2025) soulèvent le problème de l'estimation de faible probabilité pour l'alignement : « À partir d'un modèle d'apprentissage automatique et d'une distribution d'entrée formellement spécifiée, comment peut-on estimer la probabilité d'une propriété binaire de la sortie du modèle, même lorsque cette probabilité est trop faible pour être estimée par échantillonnage aléatoire? »

**Pourquoi cela est important :** Nous aimerais disposer de méthodes permettant de détecter et d'estimer la probabilité des risques extrêmes liés aux pires comportements des systèmes d'IA. Les systèmes d'IA désalignés visant à contourner la supervision humaine ne peuvent agir de manière malveillante que dans de très rares cas, où la supervision humaine est susceptible de ne pas détecter la défection. De tels cas de défection sont des événements peu probables qui ne peuvent être mesurés directement à l'aide d'un test de performances standard portant sur un petit échantillon.

#### **Exemples de sous-problèmes :**

1. Existe-t-il de nouvelles méthodes pour aborder les tests de performances dans [Wu et Hilton \(2025\)](#)?
2. Nous voulons déterminer si les systèmes d'IA contiennent suffisamment d'information pour créer des contextes dans lesquels ils pourraient se comporter de manière malveillante (le problème lié au déclenchement de contextes néfastes, [Irving et coll., 2025](#)). Peut-on extrapoler les résultats de [Wu et Hilton \(2025\)](#) pour résoudre le problème lié au déclenchement de contextes néfastes?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Wu et Hilton \(2025\)](#). Les recherches antérieures dans ce domaine comprennent [Webb et coll. \(2018\)](#), qui étudie le problème de l'estimation de faible probabilité dans le domaine de la vision par ordinateur.

### *Méthodes bayésiennes pour l'IA scientifique*

**Résumé du problème :** [Bengio et coll. \(2025\)](#) décrit une voie vers une superintelligence sécuritaire grâce à un suivi explicite de l'incertitude et à un champ d'application stratégiquement limité pour le déploiement de l'IA avancée. De manière générale, l'objectif est d'atténuer le piratage des récompenses et les comportements visant à acquérir du pouvoir en incitant l'IA à modéliser explicitement l'incertitude et à expliquer les données (y compris les textes générés par les êtres humains). Une IA scientifique

bayésienne modéliserait l'incertitude liée aux variables latentes du langage naturel, ce qui permettrait d'interroger l'IA sur ses convictions plutôt que simplement sur ce qu'un être humain dirait. Un élément central de ce programme consiste à remplacer l'apprentissage par renforcement fondé sur la recherche de modes par un échantillonnage proportionnel à la récompense, comme dans les réseaux de flot génératif; cela empêche l'effondrement des modes ([Deleu et coll., 2023](#)).

**Pourquoi cela est important :** Pour les tâches insuffisamment ou mal spécifiées, une IA agentique avancée sera incitée à exploiter les erreurs de supervision humaine, ce qui conduira à un désalignement de l'IA. Le programme d'IA scientifique lancé par Bengio et coll. aborde ce problème en proposant un nouveau cadre de modélisation des variables latentes axé sur la fiabilité des énoncés en langage naturel qui désignent ces variables.

#### **Exemples de sous-problèmes :**

##### *Sous-problèmes établis :*

1. **A priori pour l'IA scientifique bayésienne :** Quel langage est utilisé pour cet *a priori*, et comment la cohérence intra- et inter-hypothèses constitue une question centrale. Par exemple, [Mahfoud et coll. \(2025\)](#) démontre comment un réseau de flot génératif peut servir à mettre à jour un *a priori* sur les arbres de décision. Se reporter également à [Irving et coll. \(2022\)](#) pour obtenir un autre point de vue sur ce problème.
2. **Évaluation et limites des réseaux de flot génératif :** L'optimum des objectifs des réseaux de flot génératif est l'échantillonnage proportionnel à la récompense, mais nous ne savons pas comment évaluer efficacement la qualité de l'ajustement étant donné que l'optimum ne peut pas être facilement échantillé à partir de [Silva et coll. \(2025\)](#).
3. **Tests de performances relatifs à l'honnêteté, à la loyauté et à la tromperie :** La principale propriété à évaluer dans une mise en œuvre de l'IA scientifique bayésienne, outre la capacité à estimer les probabilités conditionnelles sur des énoncés en langage naturel, est sa comparaison avec les systèmes d'IA existants au regard de divers indicateurs de fiabilité des réponses.

**Pour en apprendre davantage :** Nous recommandons la lecture de [Hu et coll. \(2023\)](#) en ce qui a trait à l'application et [Bengio et coll. \(2025\)](#) pour obtenir un aperçu de l'orientation.

#### *Cadres logiques inclusifs*

**Résumé du problème :** Les modèles de pointe raisonnent en langage naturel, mais il pourrait être possible de mettre au point des modèles qui emploient un langage formel pour suivre et mettre à jour leur incertitude concernant les hypothèses. Par exemple, à partir du langage naturel, on peut améliorer le raisonnement de type « chaîne de résolution » à l'aide d'affirmations numérotées ou d'arguments typés, en formalisant les termes clés de ces affirmations au moyen de définitions ou de représentations symboliques ou en reliant les chaînes d'arguments dans des graphes acycliques orientés.

**Pourquoi cela est important :** En raison de leur ambiguïté inhérente, les déclarations en langage naturel sont sujettes à des incohérences et peuvent être utilisées délibérément pour obscurcir de l'information ou induire le public en erreur. À l'inverse, les arguments formulés dans des langages formels se prêtent plus facilement à la surveillance et à l'examen minutieux, que ce soit par des êtres humains, d'autres superviseurs ou des vérificateurs formels. En intégrant des structures formelles ou semi-formelles dans le raisonnement ou la chaîne de résolution de l'IA, et même avant qu'une interprétation logique complète soit possible, les affirmations peuvent devenir plus transparentes et plus faciles à évaluer. Cela renforcerait à son tour d'autres efforts, comme la surveillance évolutive, les tests antagonistes sous forme de débats ou des systèmes comme l'IA scientifique. Parallèlement, l'amélioration de notre capacité à passer facilement du langage naturel au langage formel pourrait élargir l'applicabilité des approches qui reposent actuellement sur une formalisation complète, comme l'[IA sécurisée](#) – un paradigme qui intègre l'IA, la modélisation mathématique et les méthodes formelles dans un flux de travail évolutif afin de produire des systèmes d'IA offrant des garanties quantitatives de sécurité dans leur contexte d'utilisation.

#### **Exemples de sous-problèmes :**

1. Quels outils ou processus interactifs peuvent soutenir la formalisation incrémentale des affirmations en langage naturel et les déductions implicites entre elles? On peut citer comme exemple une synthèse de modèles causaux ou de programmes probabilistes (tels que des graphiques de dépendance probabiliste), considérés comme des modèles mondiaux. Il existe aussi une méthode selon laquelle une IA peut réviser de manière itérative son argumentation dans une chaîne de résolution, en supprimant ou en actualisant certaines parties de l'argumentation et en les remplaçant par de nouvelles, plutôt que de régénérer une chaîne de résolution entièrement nouvelle à chaque itération ([Kim et coll., 2025](#) et [Leng et coll. 2025](#)).

2. Quels cadres logiques ou semi-logiques soutiennent la spécification dynamique et incrémentale des définitions et des types, et quelles méthodes faciliteraient leur désambiguïsation interactive?
3. Quels formalismes logiques permettent de décomposer le raisonnement de type « chaîne de résolution » en propositions atomiques tout en conservant une expressivité suffisante?
4. Peut-on concevoir des représentations hiérarchiques de modèles du monde dans lesquelles les estimations probabilistes à différents niveaux d'abstraction conservent des relations cohérentes pour l'interrogation de modèles de langage?
5. Comment la logique adjointe peut-elle fournir les fondements de cadres multimodaux qui formalisent les formes de raisonnement informel ([Pruiksma et coll., 2018](#))?
6. Quels cadres catégoriques, au-delà des évaluations de l'algèbre de Heyting, peuvent soutenir un raisonnement rigoureux dans un contexte d'incertitude logique et de non-omniscience logique?
7. L'inférence amortie neuronale peut-elle aller au-delà de la logique booléenne en tirant parti des topos bi-Heyting ou de la théorie des modèles multivalués ([Dweck, 2015](#)) pour obtenir des représentations de requêtes conditionnelles plus riches?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Dalrymple \(2024\)](#).

## Théorie de l'apprentissage

D'un point de vue scientifique, comprendre et mesurer les facteurs clés qui influencent le processus d'apprentissage, puis transformer ces facteurs en leviers de contrôle, contribuent grandement à aligner les systèmes d'IA sur des valeurs humaines complexes. Cela s'explique par le fait que l'alignement sur nos objectifs peut être compromis par des défaillances dans l'apprentissage. Par exemple, un modèle peut obtenir de bons résultats quant aux pertes pour lesquelles il a été entraîné, mais se comporter de manière dangereuse dans un contexte différent. Cela peut se produire lorsque la distribution d'entrée change, qu'un antagoniste fournit la pire requête possible, que le signal de récompense est mal spécifié, qu'une erreur humaine survient ou qu'un optimiseur caché apparaît. La théorie de l'apprentissage aide à comprendre ces points d'interruption. En fournissant des indicateurs prédictifs de défaillance, des résultats d'impossibilité et de nouvelles conceptions algorithmiques, nous pouvons formuler les problèmes d'alignement en termes plus faciles à traiter. Cette section est donc organisée autour des différentes étapes au cours desquelles un algorithme échouerait à apprendre, et de la manière dont la théorie peut prédire, détecter ou prévenir de telles défaillances.

### *Dynamique d'entraînement*

**Résumé du problème :** Les modèles fondateurs modernes sont créés à l'aide d'un processus industriel consistant à exécuter des règles de mise à jour simples pour des millions d'étapes, sur des milliards de paramètres et des trillions de segments textuels. Bien que nous ne contrôlions que quelques rotateurs, comme le choix d'optimiseur, le calendrier de taux d'apprentissage, la taille des lots, l'architecture et le curriculum de données, ces choix définissent en fin de compte un système complexe à très haute dimension dont les trajectoires déterminent ce que deviendra finalement le réseau. D'un point de vue empirique, nous observons de nombreux phénomènes comme des sauts de capacité, des lois d'échelle et des périodes critiques, mais nous ne disposons pas d'une théorie concise et prédictive qui relie ces phénomènes macroscopiques aux mises à jour microscopiques. Un programme de recherche sur la dynamique d'entraînement poserait les questions suivantes : quelles sont les variables grossières pertinentes de ce système; quels sont les attracteurs et les phases existants; comment les perturbations orientent-elles une exécution vers ceux-ci ou loin d'eux; et peut-on modéliser suffisamment bien la limite en temps continu pour anticiper où une configuration donnée va aboutir?

**Pourquoi cela est important :** Pour aligner solidement les systèmes d'IA sur les objectifs humains, il faut être capable d'orienter l'apprentissage plutôt que de simplement évaluer le point de contrôle final. Si l'on peut prévoir quand un modèle est susceptible de développer des tendances à la manipulation, ou quel calendrier de taux d'apprentissage fera basculer le modèle vers un bassin de comportements préjudiciables, on pourrait intervenir avant que des préjudices ne soient causés et avant que de l'argent ne soit dépensé. Une théorie efficace de l'entraînement nous donnerait des leviers qui s'adaptent à la puissance de calcul et fonctionnent en temps réel. En l'absence d'une telle théorie, nous construisons essentiellement l'IA générale par essais et erreurs, en espérant que l'attracteur que nous trouvons est celui que nous voulons.

### **Exemples de sous-problèmes :**

#### *Sous-problèmes établis :*

1. Comment peut-on poursuivre la mise au point de modèles de dynamique d'entraînement ou approfondir les modèles actuels ([Hu et coll., 2023](#), [Karras et coll., 2024](#), [Redman et coll., 2024](#), [Nunez et Soatto, 2024](#) et [Saxe et coll., 2019](#)) et examiner de manière critique leurs hypothèses théoriques? À partir d'un modèle de dynamique d'entraînement, peut-on prédire la sensibilité des pondérations du modèle aux données d'entraînement? Idéalement, de telles méthodes pourraient être utilisées dans le cadre d'un entraînement fondé sur l'apprentissage par renforcement afin de prévenir les modes de défaillance antagonistes ou relevant de la connivence.
2. Il existe déjà une littérature abondante sur le biais inductif de la descente de gradient stochastique, une méthode populaire pour l'entraînement de modèles par apprentissage profond. En général, quels choix en matière d'optimiseur ou d'architecture font le plus pencher le biais inductif vers un comportement utile plutôt que trompeur? Par exemple, [Lesci et coll. \(2025\)](#) propose une méthode pour estimer le biais induit par un certain choix de segmentation. Comment peut-on s'appuyer sur la théorie pour concevoir de bons biais inductifs ([Hu et coll., 2025](#))?
3. Il existe un écart considérable entre les configurations typiques d'entraînement et leurs modèles théoriques. De nombreux phénomènes empiriques et pratiques d'ingénierie n'ont pas fait l'objet d'une attention théorique. Nous souhaitons soit obtenir l'assurance raisonnable que ces lacunes dans les connaissances sont sans importance, soit acquérir une bonne compréhension théorique de celles-ci. Nous aimerais voir un catalogage rigoureux et systématique de ces mystères,

autrement dit, isoler systématiquement les variables, vérifier les hypothèses et progresser vers une compréhension théorique. Voici quelques exemples :

- a. Quel rôle joue réellement le bruit de descente de gradient stochastique en mini-lots ([Ziyin et coll., 2022](#))? Est-il essentiel à l'exploration ou s'agit-il simplement d'une commodité de calcul?
  - b. Quels sont les effets du momentum ([Xie et coll., 2022](#)) et d'autres innovations importantes pour les optimiseurs modernes?
  - c. Pourquoi certaines tailles de lots ([Shau et coll., 2024](#)) et certains taux d'apprentissage ([Jin et coll., 2023](#)) – ainsi que leurs calendriers – sont-ils plus performants que d'autres?
  - d. Qu'est-ce qui rend la normalisation par lots ([Ioffe et Szegedy, 2015](#), [Santurkar et coll., 2018](#) et [Bjork et coll., 2018](#)) et la normalisation par couches ([Ba et coll., 2016](#), [Xu et coll., 2019](#) et [Xiong et coll., 2020](#)) si efficaces, et pourquoi fonctionnent-elles dans différents contextes?
  - e. Comment l'entraînement conscient de la quantification ([Nagel et coll., 2022](#), [Park et coll., 2018](#) et [Nahshan et coll., 2018](#)) conserve-t-il sa précision malgré une discréétisation radicale des pondérations?
  - f. Pourquoi existe-t-il des activations massives (activations précises dont les valeurs sont considérablement plus élevées que les autres) ([Sun et coll., 2024](#)), et quel est leur lien avec l'encodage de position rotatif; [Jin et coll., 2025](#))?
4. Différents entraînements dotés d'architectures identiques peuvent aboutir à des solutions différentes sur le plan qualitatif : certains modèles mémorisent tandis que d'autres généralisent, certains développent des structures internes modulaires alors que d'autres restent enchevêtrés, certains font preuve d'une utilisation inoffensive des outils pendant que d'autres peuvent développer des tendances trompeuses. Ces points finaux distincts suggèrent que la dynamique d'entraînement comporte plusieurs attracteurs et phases. Peut-on déterminer systématiquement les parties importantes du « diagramme de phase » dans différentes configurations d'entraînement? Les questions clés sont les suivantes :
- a. Premièrement, quelles variables macroscopiques distinguent réellement ces phases? Par exemple, la théorie de l'apprentissage singulier suggère de suivre des mesures comme la perte d'entraînement et les coefficients d'apprentissage local ([Lau et coll., 2024](#) et [Chen et coll., 2023](#)).
  - b. Deuxièmement, quelles conditions d'entraînement (échelle d'initialisation, calendriers de taux d'apprentissage, ordonnancement des données,

- goulets d'étranglement architecturaux) permettent de sélectionner de manière fiable des attracteurs précis?
- c. Troisièmement, peut-on surveiller efficacement ces variables déterminantes pour la phase pendant l'entraînement à grande échelle, afin de pouvoir intervenir à un stade précoce?
  - d. Enfin, quels mécanismes de contrôle existent-ils et dispose-t-on d'une théorie scientifique fiable démontrant leur efficacité? Des interventions ponctuelles à des moments critiques de l'entraînement sont-elles suffisantes, ou faut-il un contrôle continu sous forme de rétroaction tout au long de l'entraînement?
5. Les réseaux de neurones biologiques présentent des périodes critiques pendant lesquelles certaines capacités doivent se développer ou être définitivement altérées ([Hubel et coll., 1970](#), [Coulson et coll., 2022](#) et [Byrne et Jerbi, 2022](#)). Les réseaux de neurones artificiels semblent afficher des propriétés semblables ([Achille et coll., 2018](#), [Kleinman et coll., 2023](#), [Nakaishi et coll., 2024](#) et [Chimoto et coll., 2024](#)). Certaines preuves suggèrent que l'entraînement précoce façonne de manière irréversible les biais inductifs ([Achille et coll., 2018](#)), mais aucune étude systématique n'a été menée à ce sujet. Peut-on déterminer quelles capacités comportent des périodes critiques et lesquelles peuvent être acquises à tout moment? L'« hypothèse du cerveau critique » issue des neurosciences, selon laquelle les réseaux ont tendance à s'autoorganiser jusqu'au seuil de transitions de phase, s'applique-t-elle aux systèmes artificiels?
6. [Kaplan et coll. \(2020\)](#) et [Hoffmann et coll. \(2022\)](#) documentent les relations de loi de puissance précises entre la perte, le calcul et la taille du modèle. Cependant, nous manquons d'explications mécanistes. Quels facteurs peuvent influencer les exposants de mise à l'échelle? Des études récentes suggèrent des liens avec la dimension des variétés de données ([Sharma et Kaplan, 2022](#)), la dynamique d'apprentissage des attributs ([Bordelon et coll., 2024](#)) et les facteurs liés à la distribution des modèles dans les données ([Michaud et coll., 2023](#)). Existe-t-il d'autres facteurs qui influencent l'apparition d'un comportement d'entraînement obéissant à une loi d'échelle? Quels détails ou hyperparamètres pertinents et non pertinents d'un algorithme d'entraînement influencent les exposants de mise à l'échelle?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. L'entraînement d'un réseau de neurones implique des milliards de paramètres, mais des descriptions utiles peuvent ne nécessiter que quelques variables macroscopiques. Que sont ces coordonnées grossières : propriétés spectrales de matrices de pondération, alignement entre les gradients et les données, dimensionnalité effective des représentations? Peut-on formuler des équations dynamiques pour ces variables qui permettent de prédire les résultats de l'entraînement? Lorsque le système est perturbé (p. ex., en raison de données corrompues, de gradients antagonistes ou de changements d'hyperparamètres), la théorie efficace peut-elle prédire si l'entraînement va se rétablir ou diverger?
2. L'apprentissage par renforcement introduit une complexité supplémentaire : la distribution des données observées par le modèle change à mesure que la politique s'améliore. Comment les attracteurs de telles dynamiques et leurs transitions se manifestent-ils dans de telles conditions? Peut-on étendre les théories efficaces issues de l'apprentissage supervisé pour gérer cette coévolution de la politique et de l'expérience?
3. Si la descente de gradient peut créer des processus d'optimisation interne, existe-t-il différentes phases avec différents objectifs de mesa ou sans aucun objectif? Peut-on détecter et contrôler le moment où nous transitons vers de tels états? Pourrait-on reconstruire des objectifs de mesa potentiels en analysant l'historique des points de contrôle, les gradients et les données d'entraînement? Quel niveau minimal de journalisation pendant l'entraînement permettrait la détection *a posteriori* d'une optimisation émergente?

**Pour en apprendre davantage :** Pour ce qui est de l'alignement et de la dynamique d'apprentissage, nous recommandons la lecture de [Park et coll. \(2024\)](#), de [Baker et coll. \(2025\)](#), ainsi que de [Piotrowski et coll. \(2025\)](#).

### *Généralisation et biais inductif*

**Résumé du problème :** Les ensembles d'entraînement ne déterminent le comportement que sur la tranche étroite d'entrées qu'ils contiennent; un nombre infini de fonctions peuvent correspondre à ces données tout en divergeant ailleurs. La fonction qu'un modèle apprend réellement est déterminée par ses biais inductifs, c'est-à-dire les préférences introduites par l'optimiseur, l'architecture, le curriculum, la segmentation, la régularisation et la configuration de l'inférence. Étant donné que le déploiement introduit inévitablement de nouvelles distributions dans le système, ce sont ces biais, plutôt que la perte empirique, qui déterminent le comportement du modèle en situation réelle. Comment chaque choix de conception façonne-t-il les biais, et comment ces biais se traduisent-ils en comportement hors distribution?

**Pourquoi cela est important :** Certains biais orientent les modèles vers un apprentissage conceptuel robuste, tandis que d'autres peuvent les orienter vers des heuristiques fragiles, des fuites de données confidentielles, la flagornerie, etc. En effet, certains biais peuvent même orienter le modèle vers des objectifs totalement trompeurs, qui restent indétectables dans l'ensemble d'entraînement. Les modèles d'avant-garde n'étant pas faciles à interpréter lors de leur déploiement, le moyen le plus pratique d'éviter une telle généralisation consiste à comprendre et à orienter ces biais.

### **Exemples de sous-problèmes :**

#### *Sous-problèmes établis :*

1. En général, quand les réseaux de neurones qui semblent alignés sur la distribution d'entraînement cessent-ils de généraliser cet alignement à de nouvelles situations, par exemple en revenant à des objectifs de substitution, en divulguant de l'information privée ou en produisant des sorties trompeuses, et comment peut-on commencer à étudier ce phénomène? Nous aimeraisons réfléchir à cette question sous l'angle des étapes du cycle de vie d'un modèle : y a-t-il des périodes critiques d'entraînement pendant lesquelles un modèle apprend la plupart de ses représentations similaires à des objectifs? Quels aspects des données d'entraînement, de l'architecture du modèle, de l'optimiseur et de la configuration de l'inférence favorisent les comportements hors distribution préjudiciables dans les réseaux de neurones? Dans quelle mesure la variance dans l'entraînement est-elle attribuable au caractère aléatoire, et à quel point les comportements de généralisation dépendent-ils du caractère aléatoire lors de l'initialisation ([Zhao et coll., 2025](#))?
2. Quelle est la probabilité que les GML développent des tendances à la tromperie ou à d'autres comportements préjudiciables? Les objectifs trompeurs ([Carlsmith](#),

[2023](#)) nécessitent probablement à la fois une généralisation de la longueur et une généralisation hors contexte ([Berglund et coll., 2023](#)), qui présentent donc également un intérêt.

3. Les modes de défaillance hors distribution dans les GML, comme la généralisation pathologique, c.-à-d. le paradoxe de l'inversion ([Zhu et coll., 2024](#)) et le débridage en plusieurs coups ([Anil et coll., 2024](#)), ne sont pas encore compris au regard des biais inductifs ou de la dynamique d'entraînement qui les provoquent. La plupart des explications actuelles se résument à des récits *a posteriori* (« le modèle s'est accroché à un raccourci fragile ») plutôt qu'à des théories quantitatives qui auraient pu prédire cette défaillance. Quels sont les paramètres les plus simples pour un jouet dans lesquels un mode de défaillance hors distribution est la généralisation la plus probable?
4. Les mesures de complexité des modèles, comme les coefficients d'apprentissage local ([Lau et coll., 2024](#), [Baker et coll., 2025](#), [Aoyagi, 2025](#) et [Hitchcock, 2024](#)), calculées pendant les phases de pré-entraînement ou de peaufinage peuvent-elles prédire la capacité en aval ou la transition des phases en matière d'honnêteté? De manière plus générale, peut-on se servir de la théorie de l'apprentissage singulier pour caractériser un « catalogue » des phases parcourues par les trajectoires d'entraînement ([Chen et Murfet, 2025](#))?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. À quelle vitesse les modèles oublient-ils les contraintes de sécurité lors d'un peaufinage continu, et peut-on fixer une borne supérieure à cette vitesse d'oubli de manière analytique?
2. Dans les contextes d'apprentissage par renforcement, peut-on définir un indicateur d'objectif trompeur (qui serait une statistique mesurable) corrélé à la présence de tout objectif interne qui s'écarte de l'objectif d'entraînement et qui est activement dissimulé par le modèle (c.-à-d. un alignement trompeur)? À quelle échelle les indicateurs d'objectifs trompeurs atteignent-ils leur pic, et des lois d'échelle simples peuvent-elles prédire l'emplacement de ce pic?
3. Nous voulons que les agents disposent de garanties fondées sur la théorie de l'apprentissage qui puissent réellement fonctionner sur du matériel d'avant-garde, afin que les preuves d'alignement s'appliquent à des systèmes réels plutôt qu'à des substituts théoriques comme les inducteurs de Solomonoff

([Kosoy, 2023](#)). Peut-on concevoir un *a priori* calculable, à la manière de l'apprentissage d'Occam, ou un langage hypothétique (p. ex., des machines à séquences [\[Kosoy, 2023\]](#), des processus de décision markoviens polytopes compositionnels [\[Kosoy, 2022\]](#)) dans lequel la planification optimale de Bayes et le regret bayésien peuvent être calculés en temps polynomial en la taille de l'entrée?

**Pour en apprendre davantage :** Au sujet de l'alignement et des biais inductifs, nous recommandons la lecture de [Park et coll. \(2024\)](#). Les recherches antérieures dans ce domaine comprennent [Betley et coll. \(2025\)](#) et [Greenblatt et coll. \(2024\)](#). Pour ce qui est de l'alignement et de la dynamique d'apprentissage, nous proposons la lecture de [Park et coll. \(2024\)](#), de [Baker et coll. \(2025\)](#), ainsi que de [Piotrowski et coll. \(2025\)](#). Pour en savoir plus sur le paradoxe de l'inversion, se reporter à [Lv et coll. \(2024\)](#). Pour plus de renseignements sur l'apprentissage hors contexte, se reporter à [Treutlein et coll. \(2024\)](#) et [Betley et coll. \(2025b\)](#). Pour plus d'information sur l'alignement trompeur, se reporter à [Hubinger et coll. \(2019\)](#).

#### *Surveillance évolutive et apprentissage de préférences*

**Résumé du problème :** On ne peut pas juger de manière fiable le comportement de l'IA dans toutes les situations auxquelles elle peut être soumise. La surveillance évolutive consiste à récompenser correctement les comportements souhaités des systèmes d'IA, même lorsque ces comportements dépassent la capacité des êtres humains à les juger efficacement. Des méthodes comme le débat ([Irving et coll., 2018](#)), la modélisation récursive des récompenses ([Leike et coll., 2018](#)) et la décomposition des tâches ([Christiano et coll., 2018](#), [Wu et coll., 2021](#)) tentent de transformer ces évaluations humaines imparfaites en une surveillance beaucoup plus rigoureuse. D'autres méthodes, comme la généralisation faible à forte ([Burns et coll., 2023](#)) et l'entraînement à la cohérence ([Wen et coll., 2025](#), [Paleka et coll., 2024](#), [Bürger et coll., 2024](#) et [Burns et coll., 2023](#)) s'affranchissent complètement du recours à des évaluations humaines. Nous avons encore besoin de preuves, tant expérimentales que théoriques, pour déterminer quand ces méthodes fonctionnent réellement, quelle est l'ampleur des erreurs restantes et si celles-ci sont susceptibles d'être exploitées par un adversaire.

**Pourquoi cela est important :** La plupart des entreprises tentent de sécuriser leurs modèles de langage en les peaufinant afin qu'ils reflètent les préférences humaines. Les systèmes de nouvelle génération, plus puissants, devront s'appuyer sur une « surveillance évolutive », qui consiste à recourir à des méthodes assistées par l'IA pour juger les réponses que l'on ne peut pas évaluer directement. Si cette surveillance venait à faire défaut subrepticement dans des domaines à enjeu élevé, tels que la science de pointe, le génie logiciel ou les travaux de laboratoire en biotechnologie, un modèle puissant pourrait causer un préjudice sans que personne ne s'aperçoive des signes avant-coureurs.

#### **Exemples de sous-problèmes :**

*Sous-problèmes établis :*

1. Si une preuve concernant une méthode de surveillance évolutive est formulée sous la forme d'un équilibre fondé sur la théorie des jeux ([Hilton et coll., 2025](#)), les propriétés de sécurité de nos garanties ne sont valables que si l'entraînement converge suffisamment ([Buhl et coll., 2025](#)). La théorie de l'apprentissage peut-elle détecter si les modèles se trouvent à un minimum local?
2. L'entraînement en ligne peut servir à réduire au minimum les variations de distribution, mais il ne fournit qu'une limite moyenne des regrets : des erreurs isolées peuvent toujours se produire. Des protocoles de surveillance évolutive comme le débat entre démonstrateur et estimateur ([Brown-Cohen et coll., 2025](#)) montrent que certaines propriétés (comme l'honnêteté) seront obtenues dans une fraction  $(1 - \epsilon)$  des cas ([Buhl et coll., 2025](#)). Existe-t-il des moyens de maintenir un taux d'erreur de  $\epsilon$  au moment du déploiement?
3. Nous aimerais pouvoir déployer des systèmes d'IA dans des contextes à enjeu élevé, où une seule erreur pourrait causer des préjudices à grande échelle ([Christiano, 2021](#)). Ces erreurs sont susceptibles de se produire en partie en raison du passage de la phase d'entraînement à la phase de déploiement. Peut-on user de techniques s'apparentant à l'entraînement en ligne lors de l'inférence, sur les actions candidates plutôt que sur les actions finales?
4. Diverses méthodes tentent d'améliorer la précision ou la sécurité des GML grâce à l'abstention ([Wen et coll., 2025](#)). Permettre aux agents de s'abstenir pourrait réduire les réponses catastrophiques, mais un recours excessif à l'abstention compromet la couverture. Peut-on appliquer cela à la création de méthodes raisonnées qui permettent aux agents de transmettre des éléments non sécuritaires ou peu fiables tout en garantissant que le superviseur apprend autant que possible et que le regret global pour l'être humain reste limité?
5. Quelques études ont été menées dans le domaine de l'autocalibration des modèles ([Kadavath et coll., 2022](#), [Steyvers et coll., 2025](#) et [Kapoor et coll., 2024](#)). Nous voulons déterminer si les systèmes d'IA contiennent suffisamment d'information pour créer des contextes dans lesquels ils pourraient se comporter de manière malveillante (le problème lié au déclenchement de contextes néfastes; [Irving et coll., 2025](#)). En théorie, une solution permettant de déclencher des contextes néfastes pourrait être utile pour générer des ensembles de données qui imitent suffisamment bien le déploiement afin de réduire au minimum le décalage dans la distribution entre les phases d'entraînement et de déploiement. Cela signifie que les méthodes d'entraînement permettant de produire des systèmes alignés sont plus susceptibles de se généraliser au moment du déploiement.

- a. Existe-t-il un paramètre de modèle « petit circuit superposé à des fonctions heuristiques » permettant d'obtenir de déclencher des contextes néfastes de manière non triviale?
- b. Peut-on se servir des « susceptibilités » de la théorie de l'apprentissage unique ([Baker et coll., 2025](#)) pour faire la lumière sur le type de modifications à apporter aux données d'entraînement afin de susciter le changement de comportement souhaité?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Les systèmes de pointe appellent déjà d'autres modèles et exécutent du code à l'interne. Il est essentiel de comprendre comment ce type de calcul autoréférent se généralise afin d'éviter les optimiseurs de mesa cachés ou les planifications trompeuses. Peut-on formaliser des agents à l'aide d'un canal de calcul interne, concevoir une procédure efficace de requête qui apprend le programme inconnu régissant ce canal, ainsi que prouver un regret épistémique ou une limite de performance?
2. Une méthode de surveillance évolutive peut-elle s'abstenir de se prononcer sur des questions non sécuritaires tout en garantissant un regret limité quant à l'utilité? Cela peut s'avérer utile si nous craignons que certains paramètres antagonistes puissent également influencer le modèle de surveillance, ou si nous souhaitons uniquement évaluer quelques-uns des nombreux résultats possibles dans ce paramètre (ce qui peut en fait être le cas par défaut pour certaines méthodes). Il faut formaliser des versions sélectives de méthodes existantes, comme le « débat sélectif » ou l'« entraînement sélectif à la cohérence ». Est-ce une méthode utile pour connaître les préférences et évaluer les sorties de l'IA?
3. Qu'apprend réellement le modèle lors des mécanismes de surveillance évolutive? Il faut surveiller la courbure des pertes et d'autres statistiques (alignement centré des noyaux des couches, rang effectif, etc.) pendant le peaufinage, et vérifier si le modèle subit des transitions brusques.
4. Dans le contexte d'un débat, des règles peuvent être établies afin que les débatteurs s'y conforment et veillent au respect de celles-ci par leurs adversaires. Une hypothèse avancée pour justifier cela est que les explications véridiques forment des variétés plus petites et plus cohérentes que les explications trompeuses. Cela suggère que lorsque le nombre de règles

logiquement cohérentes augmente, le volume de l'espace occupé par les arguments faux diminue plus rapidement que celui de l'espace occupé par les arguments vrais. Il faut formaliser cette hypothèse et la soumettre à un test empirique.

5. Peut-on déterminer un taux d'apprentissage critique pour le peaufinage du débat qui garantit la convergence vers un équilibre dans lequel l'honnêteté constitue la stratégie optimale?

**Pour en apprendre davantage :** Pour en savoir plus sur le déclenchement de contextes néfastes, nous recommandons la lecture d'[Irving et coll. \(2025\)](#) et de [Baker et coll. \(2025\)](#). Concernant le débat sur l'alignement de l'IA, nous suggérons de lire [Irving et coll. \(2018\)](#), [Brown-Cohen et coll. \(2025\)](#), ainsi que [Buhl et coll. \(2025\)](#). En ce qui a trait à la surveillance évolutive en général, nous proposons de consulter [Irving et coll. \(2023\)](#) et [Bowman et coll. \(2022\)](#).

### *Robustesse antagoniste/dans les pires cas*

**Résumé du problème :** Nous voulons que les systèmes d'IA soient déployés dans des domaines critiques comme la médecine, le droit ou l'élaboration de politiques, mais ce sont des domaines dans lesquels une seule erreur peut causer des préjudices disproportionnés. Cela vaut également pour les environnements de recherche automatisés fondés sur l'IA, où une seule erreur pourrait permettre l'autoexfiltration d'un modèle d'IA. Cela peut prendre la forme d'un seul échantillon antagoniste, d'un point de données ou gradient empoisonné ou d'une perturbation latente soigneusement choisie. Dans ces cas, on peut essayer d'éliminer les pires comportements au moyen de l'entraînement, de susciter des réactions qui rendent les pires cas impossibles, ou de réduire l'incidence de ces comportements.

**Pourquoi cela est important :** Les arguments en matière de sécurité qui reposent sur un comportement « moyen » ne fonctionnent pas si un attaquant peut forcer le modèle à entrer dans des états rares, mais non sécuritaires. Les garanties de robustesse constituent notre seule ligne de défense dès lors qu'un modèle fonctionne dans des environnements ouverts ou interagit avec des utilisatrices ou utilisateurs malveillants.

### **Exemples de sous-problèmes :**

#### *Sous-problèmes établis :*

1. L'entraînement par antagonisme peut donner lieu à une mauvaise généralisation ([Raghunathan et coll., 2019](#)) et nuire aux capacités inoffensives ([Tsipras et coll., 2019](#)), mais il est important pour éliminer les pires comportements.

L'entraînement par antagonisme assoupli ([Hubinger, 2019](#)) et l'entraînement par antagonisme latent ([Jermyn, 2022](#) et [Sheshadri et coll., 2024](#)) offrent des solutions de rechange. Peut-on mettre au point des approches empiriques qui permettent d'obtenir un taux de réussite élevé en matière de suppression de portes dérobées dont les données sont empoisonnées? Sinon, peut-on comparer le taux de succès de l'entraînement par antagonisme à celui obtenu en matière de résistance aux débridages test?

2. Des protocoles de surveillance évolutive comme le débat entre démonstrateur et estimateur ([Brown-Cohen et coll., 2025](#)) montrent que certaines propriétés (comme l'honnêteté) seront obtenues dans une fraction  $(1 - \epsilon)$  des cas ([Buhl et coll., 2025](#)). Y a-t-il des moyens de restreindre les actions que les systèmes pourront effectuer dans les cas d'erreur  $\epsilon$  restants?
3. L'entraînement à la cohérence ([Wen et coll., 2025](#), [Paleka et coll., 2024](#), [Bürger et coll., 2024](#) et [Burns et coll. 2023](#)) constitue un moyen prometteur de peaufiner les modèles de langage sans avoir recours à des préférences humaines coûteuses. L'entraînement par antagonisme s'effectue ici principalement sous forme d'étiquettes, en utilisant les représentations internes du modèle pour générer de la cohérence. Ces méthodes fonctionnent-elles encore si les pondérations elles-mêmes sont antagonistes aux valeurs humaines (comme dans le cas d'un modèle présentant un désalignement émergent; [Betley et coll., 2025](#))?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Dans quelle mesure la théorie peut-elle nous renseigner – à partir de certaines hypothèses quant à la puissance de l'adversaire – sur les préjudices les plus importants que les erreurs d'un système pourraient causer? En d'autres termes, peut-on établir des limites qui relient l'influence d'un adversaire à la gravité de la défaillance qui en résulte, allant d'erreurs insignifiantes (une faute de frappe dans une observation de jaugeage) à des conséquences catastrophiques (recommander une dose 100 fois supérieure à celle prescrite ou orienter les utilisateurs et utilisatrices vers une menace biologique)?
2. Lorsqu'un modèle reçoit pour instruction de s'autocritiquer ou de vérifier sa propre sortie à l'aide d'une chaîne de résolution, cela améliore-t-il de manière prouvable la limite de l'ensemble sécuritaire d'entrées par rapport à l'entraînement par antagonisme standard? Peut-on fixer une limite de type PAC-Bayésienne ou une limite compressive quant à la probabilité restante de débridage?

3. Si on trouve une séquence antagoniste après le déploiement, quel degré de perturbation des données ou des paramètres est nécessaire pour l'exclure de manière prouvable, en supposant qu'il n'y ait pas de réentraînement à partir de zéro?
4. Dans une attaque par infiltration de requête, un attaquant inclut une instruction (la requête) dans un texte lu par le GML (les données) dans le but d'amener le GML à suivre l'instruction infiltrée pour effectuer une tâche choisie par l'attaquant ([Shao et coll., 2024](#), entre autres). Peut-on modéliser une attaque par infiltration de requête comme une perturbation dans l'espace des attributs de la requête, et en déduire une limite du nombre de coups malveillants nécessaires pour forcer un certain taux d'erreur cible dans un GML de type GPT?

**Pour en apprendre davantage :** En ce qui concerne l'entraînement par antagonisme assoupli ou latent, nous recommandons la lecture de [Christiano \(2019\)](#), de [Hubinger \(2019\)](#) de [Casper et coll. \(2024\)](#) et de [Sheshadri et coll. \(2024\)](#). Les recherches antérieures dans ce domaine comprennent [Xhonneux et coll. \(2024\)](#). Nous suggérons de lire les études récentes suivantes consacrées au désalignement émergent : [Betley et coll. \(2025\)](#), [Soligo et coll. \(2025\)](#), ainsi que [Turner et coll. \(2025\)](#).

### *Distribution des données*

**Résumé du problème :** La base d'apprentissage constitue le principal moyen par lequel nous spécifions ce qu'un modèle doit savoir, valoriser et ignorer. Les laboratoires d'avant-garde modernes consacrent déjà des milliers d'heures de processeurs graphiques et d'énormes efforts humains au moissonnage, à la déduplication, au filtrage et à l'annotation des bases d'apprentissage de pré-entraînement. Cependant, l'objectif d'optimisation dominant reste « réduire la perte d'entropie croisée aussi rapidement que possible », et non « maintenir une compréhension précise et mécaniste de ce qu'enseigne chaque tranche de la base d'apprentissage ». Par conséquent, même si nous effectuons une curation des données, nous continuons à traiter la composition finale comme un vaste ensemble homogène, en grande partie indifférencié. De plus, nous n'avons pas encore suffisamment élaboré un langage scientifique précis pour décrire les modèles, les régularités et les attributs des données, ni mis au point d'outils évolutifs capables de détecter, d'ajouter ou de supprimer ces modèles à volonté. L'objectif est de cerner les voies d'influence entre des tranches de données précises et des comportements précis du modèle, de découvrir automatiquement les modes latents de distribution des données dans la base d'apprentissage et d'élaborer une science précise de la manière dont la structure statistique passe du texte aux pondérations.

**Pourquoi cela est important :** Les données constituent le levier le plus accessible et le plus acceptable pour façonner la cognition des modèles. Si on pouvait prédire que « l'inclusion de la base d'apprentissage X produit de façon fiable le comportement Y », on pourrait éviter les fuites de données privées, la flagornerie ou les styles indésirables. À l'inverse, on pourrait également s'assurer que les concepts essentiels à la sécurité, comme l'honnêteté, soient solidement acquis. Une telle compréhension est également une condition préalable au chargement éventuel de valeurs, afin que l'on puisse insérer de manière chirurgicale lesdits modèles dans nos ensembles de données après avoir apporté des modifications.

### **Exemples de sous-problèmes :**

#### *Sous-problèmes établis :*

1. Une fonction d'influence est une mesure statistique classique permettant d'attribuer l'effet des échantillons d'entraînement à des propriétés précises du modèle et aux résultats de l'entraînement. Ces fonctions ont été utilisées pour mieux comprendre les prédictions d'un modèle « boîte noire » ([Koh et Liang, 2017](#), [Grosse et coll., 2023](#), [Zhang et coll., 2024](#), [Kreer et coll., 2025](#), [Adam et coll., 2025](#) et [Anthropic, 2023](#), entre autres). Peut-on concevoir des méthodes évolutives pour découvrir des grappes de données significatives et attribuer de manière fiable les propriétés du modèle à ces grappes? Par exemple, peut-on

recenser le sous-ensemble minimal de données qu'il faudrait supprimer pour éliminer une capacité préjudiciable donnée ou promouvoir certaines propriétés de sécurité?

2. La base d'apprentissage est un échantillon issu d'une certaine distribution qui contient des régularités statistiques, ou modes ([Chen et coll., 2025](#) et [Saxe et coll., 2019](#)), à plusieurs échelles, allant des bigrammes de caractères aux schémas de pensée. Cependant, nous ne disposons pas encore de méthodes systématiques pour distinguer les modèles « profonds » (au sens où ils sont solidement encodés et constituent des éléments fondamentaux pour d'autres modèles, se reporter à [Lehalleur, 2025](#)) des modèles « superficiels ». Peut-on élaborer des méthodes évolutives pour repérer les modèles profonds d'une distribution de données à partir de ses échantillons ([Baker et coll., 2025](#))? De plus, en associant ces connaissances à celles relatives à l'architecture du modèle, peut-on prédire quels modes seront appris de manière fiable, dans quel ordre et à quelle échelle du modèle de données?
3. Il existe de nombreuses recherches ([Tay et coll., 2022](#), [Bahri et coll., 2021](#) et [Kaplan et coll., 2020](#), entre autres) portant sur la performance d'un système au fur et à mesure que les variables sont augmentées ou diminuées. Ces lois d'échelle sont connues pour être influencées par la modification de la composition des données (taille de l'ensemble de données utilisé pour le pré-entraînement, diversité linguistique, etc.). Pourquoi cela se produit-il et quel effet cela a-t-il sur le modèle par rapport à d'autres modifications (architecture, optimiseur, etc.)?
4. L'apprentissage par curriculum est une méthode d'apprentissage automatique qui consiste à entraîner des modèles en tenant compte de la difficulté des différentes tranches de la base d'apprentissage ([Soviany et coll., 2021](#)). Dans quelle mesure les comportements éventuels du modèle sont-ils sensibles aux permutations de l'ordre de présentation de ces tranches? De manière plus formelle, pour les tranches K, une architecture et un optimiseur donnés, quelle est la variance du comportement post-entraînement B lorsque l'ordre des données est tiré de façon aléatoire et uniforme parmi les curricula possibles de K! ? Peut-on limiter ou prédire cette « variance du curriculum » en fonction de l'échelle du modèle et des statistiques sur les tranches?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Il existe un problème inverse à la découverte de modèles et de régularités dans la distribution des données : étant donné une spécification des propriétés souhaitées pour un modèle, peut-on concevoir une distribution – ou modifier la distribution actuelle – pour obtenir ce qui est spécifié? Que faut-il pour avoir un tel « compilateur de distribution des données »?
2. Supposons qu'une base d'apprentissage pré-entraînement soit divisée en modes  $K$  identifiables (dialectes, genres de tâches, « groupes culturels », etc.). Peut-on dégager un cadre fondé sur la théorie de l'apprentissage permettant de prédire les « termes d'interaction » d'ordre supérieur qui émergent lorsque deux modes ou plus sont combinés, c.-à-d. les comportements qui sont absents lorsque les modes sont entraînés de manière isolée, mais qui apparaissent lorsqu'ils sont combinés? Quelle peut être l'ampleur de ces effets d'interaction et dans quelles conditions (tailles relatives des modes, ordre des curricula, largeur des modèles) s'amplifient-ils ou s'annulent-ils?

**Pour en apprendre davantage :**

À propos de l'effet des données sur la structure du modèle et l'apprentissage, nous recommandons la lecture de [Lehalleur \(2025\)](#). Pour en savoir plus sur les lois d'échelle neuronales, nous conseillons de lire [Michaud et coll. \(2024\)](#), [He et coll. \(2024\)](#) et [Liu et coll. \(2025\)](#).

### *Erreur de spécification et piratage des récompenses*

**Résumé du problème :** La pratique actuelle de l'alignement consiste à entraîner de grands modèles par rapport à des objectifs de substitution : modèles de récompense, classificateurs de préférences ou critères de réussite clairsemés. Ces objectifs de substitution peuvent être manipulés (le « piratage des récompenses ») ([Skalse et coll., 2022](#), [Weng, 2024](#) et [METR, 2025](#)), par exemple lorsque les systèmes d'IA tentent de tricher pour obtenir un pointage élevé sans réellement poursuivre l'objectif de leur programme. Nous devrions essayer de caractériser à quel moment l'optimisation des objectifs de substitution conduit à des politiques catastrophiques et comment prévenir ou détecter ce phénomène.

**Pourquoi cela est important :** Une politique parfaitement rigoureuse et parfaitement généralisable reste dangereuse si l'objectif qu'elle poursuit est contraire aux valeurs humaines. De plus, toute méthode d'entraînement faisant appel à des récompenses humaines ou rattachées à l'IA (apprentissage par renforcement avec rétroaction humaine, apprentissage par renforcement à partir de la rétroaction de l'IA, apprentissage par renforcement constitutionnel) héritera des faiblesses (liées au piratage des récompenses) de leurs modèles de récompense.

## Exemples de sous-problèmes :

### *Sous-problèmes établis :*

1. Peut-on mettre au point des évaluations rapides et réalistes qui signalent lorsqu'une politique a découvert une stratégie à court terme qui augmente la récompense apprise tout en réduisant une utilité mesurée indépendamment dans la réalité de terrain? Concrètement, on pourrait entraîner deux modèles de récompense sur des ensembles d'attributs disjoints; une divergence soudaine entre leurs gradients permettrait-elle de prédire un piratage des récompenses imminent?
2. En gros, la dimension de Littlestone d'une classe d'hypothèses (ou d'un apprenant limité à cette classe) est une mesure de la difficulté de la classe dans un contexte de prédiction séquentielle antagoniste. L'apprentissage en ligne ambigu ([Kosoy, 2025](#)) est un cadre récent qui permet à un apprenant de fournir plusieurs étiquettes prédites. Peut-on étendre cela aux réseaux profonds? Si oui, sa dimension de Littlestone ambiguë ([Kosoy, 2025](#)) prédit-elle mieux la généralisation hors distribution ou la résistance au débridage ([Kumar et coll., 2024](#) et [Hasan et coll., 2024](#)) que les limites de Littlestone classiques ([Shalev-Shwartz, 2011](#) et [Littlestone, 1988](#))?
3. Les récompenses cachées sont des fonctions d'utilité définies à partir d'états latents qui ne sont jamais directement observés ([Kosoy, 2025](#) et [Dogan et coll., 2023](#)). Les valeurs humaines réelles font référence à des concepts comme « la souffrance » ou « la biodiversité », et non à des flux de segments textuels. Les processus de décision markoviens partiellement observables ont des états physiques latents qui doivent être inférés, mais supposent néanmoins que la récompense est connue et renvoyée à l'agent à chaque étape. L'apprentissage par renforcement inverse, l'apprentissage par renforcement fondé sur les préférences et d'autres méthodes semblables permettent certes d'apprendre un modèle de récompense, mais ils le fondent sur le vecteur d'attributs ou sur l'historique des observations plutôt que de chercher à spécifier une variable réelle. Certaines études ([Ha et Schmidhuber, 2018](#), [Richens et Everitt, 2024](#) et [Kipf et coll., 2020](#)) tentent d'apprendre des variables latentes censées correspondre à des « objets » ou à des « causes physiques », puis d'apprendre des politiques qui optimisent une récompense définie à partir de ces variables latentes. Mais ce qui constitue l'espace latent approprié est laissé à la discréption de l'apprenant non supervisé; il n'est pas nécessaire qu'il corresponde à des concepts conformes à l'éthique ou porteurs de valeurs.

- a. Nous devrions essayer de spécifier et d'apprendre ces objectifs à l'état latent, car cela semble important pour un alignement externe non trivial. Peut-on les formaliser et prouver les résultats d'assimilabilité ou d'impossibilité?
- b. Une formulation possible serait la suivante : supposons que le monde soit un processus de décision markovien partiellement observable avec un état  $S$  non observé. Il existe une fonction de récompense réelle  $R^* : S \rightarrow R$  qui encode ce qui importe aux êtres humains. L'agent ne reçoit que des observations  $o_t$ , choisit des actions  $a_t$  et reçoit peut-être des étiquettes très parcimonieuses  $L(o_{\{1:t\}}, a_{\{1:t\}})$  comme des comparaisons de préférences ou des pointages scalaires occasionnels. Nous poserions ensuite les questions suivantes :
  - i. Peut-on noter les conditions dans lesquelles  $R^*$  est (ou n'est pas) identifiable à partir des données dont l'agent peut disposer?
  - ii. Étant donné l'identifiabilité, peut-on limiter la complexité échantillonnale de l'apprentissage de  $R^*$  à une certaine plage de précision dans une norme donnée?
  - iii. Si l'on ne peut pas déterminer  $R^*$  avec exactitude, peut-on au moins obtenir une enveloppe prudente  $R \leq R^* \leq \bar{R}$  suffisamment restrictive pour exclure toute action catastrophique?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. L'alignement s'effondre si une action catastrophique ou une petite erreur de spécification du modèle annule les garanties de sécurité; une limite de regret agnostique et conscient des pièges ([Kosoy, 2023](#)) constituerait la première réponse formelle à ce mode de défaillance. Peut-on produire un algorithme dont le regret reste sous-linéaire même lorsque i) l'environnement contient des états « pièges » irréversibles et ii) l'environnement réel peut se situer en dehors de la classe d'hypothèses (les jeux de devinette de mot de passe constituerait le banc d'essai canonique)?
2. Une mise à jour « améliorant la valeur » correspond à toute orientation du gradient de politique qui augmente l'utilité réelle (latente) qui nous intéresse, tandis qu'une mise à jour de type Goodhart correspond à une orientation qui augmente la récompense substitutive apprise sans augmenter l'utilité réelle.

Pour une chaîne de traitement d'apprentissage par renforcement avec rétroaction humaine donnée (comme la configuration de [Gao, 2024](#)), peut-on décomposer le champ vectoriel des gradients de politique en composantes « améliorant la valeur » et en composantes de type Goodhart, puis fixer une borne supérieure pour la divergence à long terme?

3. Peut-on définir les conditions nécessaires et suffisantes pour que le modèle de récompense à état latent appris soit calibré par rapport à une récompense cachée de réalité de terrain, en supposant une observabilité partielle et des *a priori* mal spécifiés?
4. Les modèles peaufinés choisissent de plus en plus leurs propres données d'entraînement (jeu contre soi-même, collecte automatisée de données). Peut-on créer des tests en ligne capables de détecter lorsqu'un apprenant renforce une boucle de piratage des récompenses plus rapidement que la surveillance humaine ne peut réagir?

**Pour en apprendre davantage :** Au sujet de la loi de Goodhart et des jeux de spécification, nous suggérons de lire [Manheim et Garrabrant \(2018\)](#) et [Krueger et coll. \(2020\)](#). En ce qui concerne le piratage des récompenses, nous recommandons la lecture de [METR \(2025\)](#), de [Weng \(2024\)](#), de [Hadfield-Menell et coll. \(2016\)](#) et d'[Everitt et coll. \(2019\)](#). Les recherches antérieures dans ce domaine comprennent [Wen et coll. \(2024\)](#), [Pan et coll. \(2024\)](#), ainsi que [Wang et coll. \(2024\)](#).

### *Apprentissage contextuel et continu*

**Résumé du problème :** Les grands modèles apprennent de plus en plus au moment de l'inférence, que ce soit en absorbant de nouvelles requêtes informatives, en misant sur l'utilisation d'outils ou en intégrant des données transmises en continu. Ces processus d'apprentissage peuvent être très différents, mais ils sont considérablement influencés par l'entraînement préalable, y compris le pré-entraînement et le post-entraînement. Par ailleurs, ils progressent quant à l'horizon temporel dans lequel ils opèrent, la mémoire à long terme, ainsi que l'évolution de leurs objectifs et valeurs. Cela peut entraîner des défis techniques, notamment pour ce qui est d'éviter l'oubli catastrophique des contraintes de sécurité ou de détecter le moment où de nouvelles capacités ou valeurs sont acquises. Nous aimerais disposer d'une théorie ou d'outils de mesure permettant de quantifier la vitesse à laquelle les connaissances en matière de sécurité se dégradent, les choix d'architecture ou d'optimiseur qui influencent cette dégradation, et comment des optimiseurs de mesas pourraient apparaître dans un tel contexte.

**Pourquoi cela est important :** Les garanties de sécurité obtenues lors de l'entraînement initial ne servent à rien si elles disparaissent après une semaine d'utilisation. Plusieurs comportements toxiques n'apparaissent qu'après de longues périodes d'interaction avec un utilisateur ou une utilisatrice. Des mécanismes d'oubli fiables, une adaptation en ligne sécuritaire et des indicateurs d'alerte précoce en cas de comportement préjudiciable sont donc essentiels pour maintenir l'alignement des systèmes déployés tout au long de leur cycle de vie. Sans eux, chaque peaufinage ou fenêtre contextuelle post-déploiement devient un vecteur potentiel de débridage, érodant progressivement les mesures de sécurité initialement mises en place.

### **Exemples de sous-problèmes :**

#### *Sous-problèmes établis :*

1. L'apprentissage contextuel désigne la capacité émergente d'un modèle à ajuster son comportement en fonction des segments textuels qu'il reçoit pendant l'inférence et à améliorer efficacement ses capacités existantes ou à en acquérir de nouvelles ([Brown et coll., 2020](#) et [Dong et coll., 2022](#)). Nous avons une compréhension approximative du moment où les modèles acquièrent et perdent leur capacité d'apprentissage contextuel ([Singh et coll., 2023](#) et [Carroll et coll., 2025](#)), ainsi que des conditions de pré-entraînement qui favorisent l'apprentissage contextuel. Nous ne comprenons encore pas certaines choses, comme la manière de contrôler les caractéristiques de l'apprentissage contextuel par l'entremise d'interventions avant, après et pendant l'apprentissage par

renforcement, ni la théorie efficace qui explique le lien entre les capacités d'apprentissage contextuel et la dynamique pré-entraînement ([Singh et coll. 2024](#)).

2. Divers mécanismes ont été proposés pour expliquer l'apprentissage contextuel. Peut-on valider ou réfuter les mécanismes suivants?
  - a. L'apprentissage contextuel comme optimiseur de mésa à descente de gradient dans la propagation avant : [Oswald et coll. \(2023\)](#)
  - b. L'identification des têtes d'induction comme composants essentiels et centraux de la capacité d'apprentissage contextuel : [Olsson et coll. \(2022\)](#)
  - c. Différents types d'algorithmes d'apprentissage mis en œuvre de manière contextuelle selon différents facteurs : [Akyürek et coll. \(2023\)](#)
3. L'oubli catastrophique ([De Lange et coll., 2021](#) et [Wang et coll., 2025](#)) désigne la perte de capacités précédemment acquises après l'entraînement sur de nouvelles données. Cela peut être préjudiciable lorsque l'on oublie les règles de sécurité apprises au cours de l'entraînement, mais peut s'avérer bénéfique lorsque l'on désapprend les comportements préjudiciables acquis pendant l'entraînement ([Yao et coll., 2024](#)). Pourquoi cela se produit-il, et comment peut-on cerner et contrôler les capacités qui sont oubliées? Existe-t-il des méthodes d'entraînement continu (apprentissage par renforcement, peaufinage ou pré-entraînement continu) et de curation minutieuse de données contextuelles qui permettent de conserver les propriétés de sécurité? Quelles mesures (tests de performances, méthodes « boîte blanche » ou autres indicateurs) peut-on effectuer pour s'assurer que les propriétés de sécurité sont conservées?
4. Le déclenchement de contextes néfastes à partir de modèles ([Irving et coll., 2025](#)) peut être lié aux recherches visant à déterminer si les modèles de langage savent ce qu'ils savent ([Kadavath et coll., 2022](#), [Steyvers et coll., 2025](#) et [Kapoor et coll., 2024](#)). Pourrait-on appliquer ces méthodes pour dégager des scénarios dans lesquels un modèle donné fonctionnerait mal ou agirait de manière préjudiciable? Peut-on mettre au point des dispositifs de surveillance de chaînes de résolution capables d'alerter rapidement en cas de transition potentielle vers des contextes dangereux? Peut-on découvrir systématiquement les requêtes potentielles de débridage de modèles grâce à une compréhension suffisante de la manière dont les modèles actualisent leurs états en contexte?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Peut-on trouver des moyens évolutifs pour surveiller la dynamique de l'apprentissage après le pré-entraînement et servir de systèmes d'alerte précoce pour tout changement de phase potentiel?
2. Pourrait-on recourir à la théorie de l'apprentissage pour comprendre l'efficacité des échantillons de divers apprenants contextuels, ainsi que leur capacité à généraliser et d'autres comportements hors distribution?
3. Quelles sont les propriétés de sécurité pour l'apprentissage continu ou pour les systèmes d'IA qui sont soit entraînés avec une mémoire à long terme, soit enrichis par une telle mémoire? Quelle est la probabilité qu'ils se comportent de manière répréhensible ou erratique, en particulier après de longues périodes d'interaction avec l'utilisateur ou l'utilisatrice?

**Pour en apprendre davantage :**

### *Alignment interne et fondements de l'agent*

**Résumé du problème :** La sélection naturelle a « entraîné » les êtres humains à maximiser leur aptitude génétique, mais elle l'a fait en nous conditionnant à rechercher des récompenses substitutives faciles à obtenir, comme le goût sucré. Dans l'environnement ancestral, ces substituts étaient synonymes de survie, mais dans le monde moderne, ce même conditionnement peut nous pousser vers les boissons gazeuses et d'autres aliments malsains. Lorsque nous entraînons un réseau de neurones, nous craignons que certaines parties du modèle commencent à agir comme leur propre optimiseur, en effectuant une recherche interne pour atteindre un objectif qui n'est pas celui que nous avions prévu. Ce phénomène a été observé dans des réseaux de neurones en configurations restreintes ([Guez et coll., 2019](#) et [Taufeeque et coll., 2024](#)). Nous devons déterminer si ces objectifs cachés apparaissent, à quel moment ils apparaissent, comment ils évoluent pendant l'entraînement et comment les détecter avant qu'ils ne causent des problèmes. En général, nous avons également besoin d'un cadre normatif : un moyen d'énoncer ce qu'un agent doit croire, vouloir et faire, ainsi que des preuves que ces prescriptions restent valables après l'apprentissage, l'autoamélioration ou l'interaction avec d'autres agents.

**Pourquoi cela est important :** Un modèle qui passe discrètement à un objectif différent peut compromettre même l'objectif d'entraînement le mieux conçu. Ces changements de comportement peuvent survenir soudainement et sans signe avant-coureur apparent, ou peuvent être déclenchés par un environnement inadapté à ces pulsions. Comprendre la dynamique d'apprentissage qui les sous-tend nous permettrait de repérer et de corriger les problèmes d'alignement avant qu'ils ne prennent de l'ampleur. De plus, sans base solide, on se retrouve avec des correctifs ponctuels dont la sécurité disparaît dès que le système réécrit ses propres règles de décision ou fait face à des changements de distribution. Une théorie décisionnelle calculable et consciente du processus d'apprentissage nous permettrait d'aborder de manière raisonnée le piratage des récompenses, l'alignement trompeur ou la connivence, ainsi qu'éclairerait la conception d'algorithmes d'entraînement dont les objectifs survivent à la mise à l'échelle. En bref, des fondements clairs pour les agents permettent de passer de conjectures empiriques à quelque chose que l'on peut prouver (ou réfuter) avant le déploiement.

### **Exemples de sous-problèmes :**

#### *Sous-problèmes établis :*

1. [Hu et coll. \(2023\)](#) suggère que les résultats de l'entraînement peuvent être visibles dès le début du processus d'entraînement. Quand le comportement d'un

modèle peut-il être interprété uniquement à partir des pondérations entraînées, et quand la dynamique d'apprentissage est-elle nécessaire à l'interprétabilité (se reporter à [Lehalleur, 2025](#) pour plus de contexte)?

2. Le déploiement ouvert et hétérogène mettra en concurrence de nombreuses IA entre elles et contre les êtres humains (se reporter à certaines études sur l'intelligence stratégique dans les GML par [Payne et Alloui-Cros, 2025](#)). Cela garantit que la coopération (ou au moins des dommages limités) émergera même dans le cadre d'un jeu croisé antagoniste, ce qui constitue un élément clé de la robustesse dans le monde réel. Les systèmes optimisés à cet effet peuvent également apprendre des dynamiques préjudiciables dans un environnement concurrentiel. Peut-on prouver qu'une classe d'algorithmes d'apprentissage (p. ex., les programmes d'apprentissage infrabayésiens ou à descente de miroir) converge vers des équilibres logit ou  $\epsilon$ -Nash (un ensemble de stratégies dans lequel, étant donné les stratégies des adversaires, chaque joueur obtient un gain maximal possible à moins de  $\epsilon$  près) dans des jeux de population répétés où les adversaires sont remaniés à chaque tour? Peut-on quantifier le pire manque à gagner possible?
3. Une dynamique cachée « agent à l'intérieur de l'agent » pourrait conduire à une connivence secrète ([Motwani et coll., 2024](#)) ou à un piratage de gradient ([Barnett, 2021](#)). Dans un tel cas, un modèle pourrait modifier son propre processus d'entraînement. Une chaîne de traitement de détection et d'atténuation offre un dernier filet de sécurité contre les pires scénarios avant le déploiement. Pourrait-on faire appel au formalisme infrabayésien de la « loi des clones » (Direction 15 dans [Kosoy, 2023](#)) pour détecter quand deux sous-réseaux d'un même modèle commencent à se modéliser mutuellement en tant qu'agents ayant des objectifs privés? Serait-il possible de concevoir une intervention (régulateur de pondération, modification architecturale, etc.) qui permettrait de supprimer cela de manière prouvable? Un bon banc d'essai serait de recourir à cette intervention pour prévenir le désalignement émergent chez des organismes modèles chez lesquels nous savons qu'il peut être induit ([Turner et coll., 2025](#)).

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

3. Les modèles actuels fondés sur la théorie de la décision (AIXI [[Hutter, 2000](#)], apprentissage par renforcement bayésien classique, théorie des jeux standard) supposent soit une puissance de calcul illimitée, soit des récompenses parfaitement observables, soit une distinction claire entre l'agent et l'environnement. Aucune de ces hypothèses ne s'applique aux systèmes d'avant-garde qui sont intégrés dans le monde, peuvent modifier leur propre code ou leur environnement et peuvent interagir avec d'autres IA. Cependant, nous avons besoin de fondements qui définissent ce qu'un agent doit croire, vouloir et faire, ainsi que des preuves que ces prescriptions restent valables après l'apprentissage, l'autoamélioration ou l'interaction avec d'autres agents. Nous posons donc les questions suivantes : Quel est le langage formel approprié pour raisonner à propos d'agents d'apprentissage limités qui doivent inférer simultanément comment fonctionne le monde et comment il convient d'agir dans celui-ci? Comment doit-on créer des modèles qui reflètent le biais de simplicité sans générer de concepts préjudiciables? Peut-on déduire des garanties (limites de regret, concepts d'équilibre, contraintes de sécurité) qui restent valables lorsque l'agent fait partie de son environnement et peut être modifié par celui-ci?
  - a. Est-il possible de formaliser les biais de simplicité au-delà de Kolmogorov, Solomonoff et Chaitin, puis de les relier aux biais inductifs au sein de modèles statistiques calculables, comme les réseaux de neurones artificiels?
  - b. Pourrait-on s'appuyer sur la théorie de l'apprentissage pour clarifier formellement la formation des optimiseurs de mesa et la définition des objectifs, ainsi que pour préciser les conditions théoriques dans lesquelles elles apparaissent?
  - c. Pourrait-on miser sur la théorie de l'apprentissage pour fournir une caractérisation formelle et quantitative du piratage des récompenses et d'autres comportements de type « généralisation erronée des objectifs » chez les agents?
4. Nous voulons une approche empirique de la dérive de valeur induite par le changement de représentation, où des modèles volumineux et continuellement peaufinés sont confrontés à une ontologie radicalement différente lors de leur utilisation. Peut-on concevoir une « attaque par mélange d'ontologies », dans laquelle on passe d'une politique entraînée sur une ontologie à une politique plus compressée lors du déploiement?

5. Les schémas d’alignement actuels supposent tacitement que les canaux d’entrée et de sortie de l’agent sont les seuls éléments qui comptent. Si des modèles avancés font disparaître cette limite (p. ex., en se simulant eux-mêmes ou en simulant leurs opérateurs), nous avons besoin de garanties qui survivent au changement ontologique. Serait-il possible de produire un modèle intégré ([Demski et Garrabrant, 2019](#) et [Kosoy et Appel, 2021](#)) qui évite les pièges classiques liés à l’autostimulation du circuit de récompense?

**Pour en apprendre davantage :** En ce qui concerne l’alignement interne, nous recommandons la lecture de [Hubinger et coll. \(2019\)](#) et de [Tennant et coll. \(2025\)](#). Au sujet de l’infrabayésianisme, nous suggérons de lire [Kosoy et Appel \(2021\)](#). Quant aux fondements des agents, nous suggérons la lecture d’ouvrages antérieurs comme UAI/AIXI de Marcus Hutter et Shane Legg ([Legg et Hutter, 2006](#) et [Hutter, 2000](#)), l’*a priori* sur la vitesse de Jürgen Schmidhuber ([Schmidhuber, 2002](#)), le programme fondé sur la théorie de l’apprentissage de Vanessa Kosoy ([Kosoy, 2023](#)) et l’induction logique pour la prise de décision dans un contexte d’incertitude logique de Scott Garrabrant ([Garrabrant et coll., 2016](#)).

## Évaluation et garanties dans l'apprentissage par renforcement

### *Évaluations de l'alignement antagoniste et organismes modèles*

**Résumé du problème :** Nous aimerais pouvoir mettre à l'essai nos méthodes d'alignement. Dans le cadre d'une approche empirique, nous pouvons créer des systèmes délibérément désalignés et voir si nos méthodes permettent de corriger, au moyen de l'entraînement, les comportements désalignés. Les modèles entraînés de manière antagoniste sont appelés « organismes modèles de désalignement » ([Hubinger et coll., 2023](#)). Nous souhaitons soumettre les méthodes d'alignement à des tests de résistance sur un ensemble d'organismes modèles aussi diversifié que possible. Des organismes modèles désalignés peuvent être construits à l'aide de l'une des techniques suivantes : la formulation de requêtes antagonistes, l'orientation de l'activation, l'apprentissage par renforcement, l'empoisonnement de données et les environnements antagonistes.

**Pourquoi cela est important :** Ce domaine de recherche pourrait fournir des preuves utiles quant à l'efficacité des méthodes d'alignement. Des comparaisons scientifiques *entre différents dispositifs d'évaluation et organismes modèles* permettront d'orienter les choix méthodologiques vers de meilleures pratiques pour l'étude empirique des méthodes d'alignement.

#### **Exemples de sous-problèmes :**

##### *Sous-problèmes établis :*

1. Comment peut-on élaborer un modèle comportant des portes dérobées intentionnelles, idéalement résistantes à l'entraînement par antagonisme standard et aussi réalistes que possible? Une équipe rouge disposant d'un accès illimité aux pondérations peut-elle trouver la porte dérobée? À quel point la porte dérobée peut-elle être rendue résistante?
2. Il a été démontré que l'entraînement par antagonisme peut, dans certains cas, agraver les portes dérobées ([Hubinger et coll., 2024](#)). Ce résultat est-il reproductible, et dans quelle mesure peut-il être généralisé?
3. Le piratage des récompenses dans les environnements d'encodage constitue un problème bien documenté dans les modèles d'avant-garde ([METR, 2025](#); [Bondarenko et coll., 2025](#) et [Baker et coll., 2025](#)). Baker et coll. montre que certains protocoles de surveillance des processus induisent des piratages des récompenses de plus en plus difficiles à détecter. Nous souhaitons mettre au point des chaînes de traitement pour des organismes modèles pouvant servir à

des évaluations antagonistes et à des bancs d'essai empiriques en matière de piratage des récompenses.

4. [Betley et coll. \(2025\)](#) constate que le peaufinage des modèles à partir d'ensembles de données restreints sur des comportements indésirables (p. ex., des codes truffés de bogues) entraîne un désalignement généralisé dans d'autres domaines, qu'ils qualifient de « désalignement émergent ». [Turner et coll. \(2025\)](#) a reproduit ce résultat sur plusieurs ensembles de données et modèles afin de créer une série d'organismes modèles pour le désalignement émergent. Nous nous intéressons à d'autres contextes dans lesquels un désalignement émergent peut se produire, ainsi qu'à la question de savoir si l'apprentissage par renforcement dans des environnements imparfaits (se reporter à la section « Erreur de spécification et piratage des récompenses » ci-dessus) induit un désalignement.

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Proposer de nouveaux protocoles pour la création d'« organismes modèles » qui reposent sur des hypothèses différentes concernant l'origine des désalignements et la manière dont les objectifs désalignés sont définis.
2. Peut-on écrire une politique « cheval de Troie » dans un environnement d'apprentissage par renforcement dont le comportement répréhensible n'est activé qu'après k étapes ou k unités de récompense afin d'émuler une tromperie à long terme? Pourrait-on entraîner un modèle à faire la même chose?
3. Parfois, et peut-être même lors d'un déploiement réel, nous trouverons un signal ambigu de désalignement (disons que le modèle prend une mauvaise décision, mais que cela pourrait être attribué à un malentendu). A-t-on un moyen d'*amplifier cette preuve*, en menant des expériences plus exhaustives sur des variations de ce même point de données, afin de déterminer clairement s'il s'agit d'un exemple de désalignement grave? On pourrait vérifier si c'est facile à faire sur différents organismes modèles.
4. Certaines des instructions ci-dessus nécessitent d'avoir une idée raisonnable du moment où un point de données devient trop antagoniste. Par exemple, lorsque nous avons simplement débridé le modèle, ce qui ne doit pas être interprété comme un désalignement inhérent. Comment quantifier le niveau d'antagonisme? Quelles propriétés nos évaluations de l'alignement doivent-elles

posséder pour garantir qu'elles repèrent les comportements répréhensibles généralisés, cohérents et ciblés, plutôt que des débridages arbitraires?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Hubinger et coll. \(2024\)](#) et de [Marks et coll. \(2025\)](#). Les recherches antérieures dans ce domaine comprennent [Greenblatt et coll. \(2024\)](#).

### *Examen empirique des protocoles de surveillance évolutive*

**Résumé du problème :** Pour réduire la supervision des problèmes difficiles à la supervision des problèmes traitables par l'être humain, les méthodes de surveillance évolutive introduisent soit des jeux antagonistes, comme dans la sécurité de l'IA par le débat ([Irving et coll., 2018](#)) et les jeux entre démonstrateur et vérificateur ([Kirchner et coll., 2024](#), soit des objectifs non supervisés, comme dans l'entraînement à la cohérence ([Wen et coll., 2025](#)). Peut-on examiner de manière empirique l'efficacité de ces protocoles?

**Pourquoi cela est important :** La surveillance évolutive consiste à récompenser correctement les comportements souhaités des systèmes d'IA, même lorsque ces comportements dépassent la capacité des êtres humains à les juger *efficacement*.

### **Exemples de sous-problèmes :**

#### *Sous-problèmes établis :*

1. Des expériences empiriques d'apprentissage automatique par le débat, comme [Khan et coll. \(2024\)](#) et [Kenton et coll. \(2024\)](#) ont eu recours à des protocoles semblables à la méthode originale du débat récursif, et non aux variantes probabilistes explorées dans [Brown-Cohen et Irving \(2024\)](#) ou dans [Brown-Cohen et coll. \(2025\)](#). Peut-on mettre en œuvre l'un de ces protocoles probabilistes avec les GML et les ensembles de données actuels (p. ex., GPQA), puis mesurer son efficacité?
2. Les techniques de surveillance évolutive nécessitent souvent des modèles très volumineux pour trouver des signaux empiriques. Il s'est avéré difficile de démontrer des résultats positifs au-delà d'une interaction minimale de deux tours, même dans des modèles d'avant-garde ([Khan et coll., 2024](#) et [Kenton et coll., 2024](#)). Peut-on créer des paramètres empiriques qui reflètent les caractéristiques d'une surveillance évolutive, c.-à-d. des signaux de préférence faibles pouvant être amplifiés pour superviser un modèle performant, mais où le modèle étudiant reste limité? De tels paramètres permettraient de mener des études empiriques à plus petite échelle, ce qui accélérerait la recherche.

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Les protocoles de surveillance évolutive sont tributaires de la stabilité (à savoir que de légères modifications des entrées ne doivent pas générer des sorties radicalement différentes). Nous pourrions ne pas être en mesure de prouver la stabilité de nos algorithmes. Il faut mener des expériences empiriques afin d'examiner les effets de petites perturbations sur les protocoles de surveillance évolutive dans la pratique.

**Pour en apprendre davantage :** Nous recommandons la lecture de [Buhl et coll. \(2025\)](#) et de [Brown-Cohen et coll. \(2025\)](#). Les recherches théoriques antérieures dans ce domaine comprennent [Irving et coll. \(2018\)](#), [Barnes \(2020\)](#), [Brown-Cohen et coll. \(2023\)](#), ainsi que [Brown-Cohen et Irving \(2024\)](#); les recherches empiriques précédentes comprennent [Khan et coll. \(2024\)](#) et [Kenton et coll. \(2024\)](#).

*Méthodes de détection et de prévention du sandbagging et du bidouillage exploratoire*

**Résumé du problème :** Le bidouillage exploratoire survient lorsqu'un agent refuse stratégiquement de donner sa meilleure réponse afin d'empêcher que la poursuite de l'entraînement modifie son comportement ([Benton et coll.. 2024](#)). Il s'agit d'une forme précise de *sandbagging*, qui peut survenir au cours d'un apprentissage par renforcement. Le bidouillage exploratoire existe-t-il dans la pratique? Peut-on détecter et prévenir ce comportement?

**Pourquoi cela est important :** Si les agents sont capables d'effectuer du bidouillage exploratoire, cela pourrait signifier qu'un entraînement pratique ne suffirait pas à faire ressortir toutes les capacités du modèle. Pour les protocoles de supervision évolutive comme le débat, cela pourrait signifier qu'un débateur est capable de trouver un contre-argument qui démontre qu'un argument est erroné, mais évite de le faire, ce qui aboutit à une conclusion incorrecte ([Hilton et coll.. 2025](#)). Le bidouillage exploratoire offre ainsi des marges de manœuvre qui pourraient être utilisées de manière antagoniste par les modèles pour causer un préjudice, et ce, même si l'objectif de l'entraînement correspond à l'idée que toutes les politiques optimales en termes de récompense sont sécuritaires (et ne constituent pas du *sandbagging*).

**Exemples de sous-problèmes :**

1. On pourrait entraîner délibérément des modèles à se livrer à du bidouillage exploratoire, créant ainsi un « organisme modèle » à cet égard ([Hubinger et coll., 2024](#)). Un tel organisme modèle peut ensuite permettre d'étudier les moyens de prévenir le bidouillage exploratoire.

2. Existe-t-il des méthodes de régularisation (prime d'entropie, contrainte KL, ré-échantillonnage avec remise pessimiste) dont on peut se servir pour empêcher le bidouillage exploratoire sans trop nuire à l'efficacité de l'échantillonnage?
3. Peut-on mesurer l'ampleur du bidouillage exploratoire? Une approche consisterait à considérer l'emploi d'un échantillonnage de type « best-of-n » d'un modèle digne de confiance pour estimer les possibilités de bidouillage exploratoire compte tenu d'un calcul d'inférence plus important. Cela pourrait ensuite servir à estimer une limite inférieure de l'exploration qu'un modèle devrait être capable d'effectuer avec davantage de puissance de calcul pour l'entraînement. De manière semblable, nous pourrions amortir le coût de cette estimation en entraînant un modèle prédictif d'exploration afin de déterminer les rendements de l'exploration, puis l'utiliser pour estimer si un modèle a épuisé les options d'exploration à rendement élevé.
4. Le bidouillage exploratoire peut-il être exploité par un adversaire plus habile? Une approche consisterait à recourir à l'apprentissage par renforcement antagoniste pour entraîner davantage un joueur tout en maintenant la politique de l'autre joueur constante. Cela permettrait de détecter le bidouillage exploratoire : si la récompense du joueur constant ne diminue pas de manière considérable, cela prouve qu'il n'y a pas eu de bidouillage exploratoire.

**Pour en apprendre davantage :** Les recherches antérieures dans ce domaine comprennent [Greenblatt et coll. \(2024\)](#), [Roger et Shlegeris \(2023\)](#), [Shlegeris et Stastny \(2025\)](#), ainsi que [Gasteiger et coll. \(2025\)](#). Concernant l'exploration dans l'apprentissage par renforcement, nous recommandons la lecture de [Hazan et coll. \(2019\)](#) pour connaître la théorie pertinente et de [Madan et coll. \(2025\)](#) pour découvrir les méthodes récentes. D'autres études pertinentes comprennent [Liu et Lai \(2023\)](#).

*Approches relatives à la recherche inexploitable fondées sur l'apprentissage par renforcement*

**Résumé du problème :** La surveillance évolutive vise généralement à fournir des garanties quant aux contextes dans lesquels les défaillances individuelles ne sont pas catastrophiques. Cependant, si votre fonction de récompense admet plusieurs solutions, vous devez également éviter que ces paramètres libres soient exploités au fil du temps (le « problème de la recherche exploitable »; [Pfau et Irving, 2025](#)). Les méthodes de recherche inexploitables devraient contraindre un modèle à répartir aléatoirement ses performances entre tous les objectifs secondaires antagonistes auxquels il a accès – dans un sens qui serait formalisé par la méthode.

**Pourquoi cela est important :** Nous aimerais pouvoir confier à l'IA des tâches insuffisamment spécifiées (p. ex., l'encodage et le conseil en recherche), pour lesquelles il existe de nombreuses solutions correctes. Cependant, une IA désalignée pourrait sélectionner de manière antagoniste des solutions correctes, mais néanmoins préjudiciables, comme l'ouverture d'une surface pour de futures exploitations ou la conservation d'information stéganographique en vue d'un comportement répréhensible ultérieur.

### **Exemples de sous-problèmes :**

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Peut-on élaborer un protocole « équipe rouge/équipe bleue » pour mettre à l'essai des solutions au problème de recherche exploitables?
2. Est-il possible de tester si des méthodes comme les réseaux de flot génératif, l'apprentissage par renforcement à entropie maximale, l'apprentissage par renforcement régularisé par KL et les jeux de diversité peuvent être utilisées pour la recherche inexploitables?
3. Peut-on mettre en place une boucle « équipe rouge/équipe bleue » dans laquelle l'équipe bleue s'entraîne à maximiser les pires pertes possibles pour l'inspecteur et l'équipe rouge corrige l'inspecteur de manière itérative? Quelle serait l'efficacité de cette boucle?
4. Peut-on écrire une tâche simple de test de performances pour révéler les défaillances de recherche exploitables?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Pfau et Irving \(2025\)](#). Des études connexes sur la diversité dans l'apprentissage par renforcement ont notamment porté sur la classe de méthodes de réseaux de flot génératif relatives à l'échantillonnage proportionnel à la récompense ([Bengio et coll., 2023](#)), ainsi que sur les contraintes latentes de distribution dans l'apprentissage par renforcement ([Lee et coll., 2025](#)).

### *Essais expérimentaux d'alignement trompeur*

**Résumé du problème :** Dans l'IA avancée, l'entraînement à l'alignement peut échouer lorsqu'un modèle parvient à distinguer les données d'entraînement des données de test et à exploiter cette connaissance pour se conformer de manière sélective aux objectifs d'entraînement. Ce phénomène est appelé « alignement trompeur ». L'alignement trompeur est un type particulier de défaillance hors distribution qui correspond au pire scénario possible. De manière empirique, [Greenblatt et coll. \(2024\)](#) démontre un

exemple de ce phénomène appelé *falsification d'alignement* dans les GML : « se conformer de manière sélective à son objectif d'entraînement afin d'éviter toute modification de son comportement hors entraînement ». [Hubinger et coll. \(2024\)](#) étudie l'entraînement et l'atténuation des *agents dormants* – un modèle particulier de désalignement dans lequel les modèles sont entraînés à être désalignés de manière sélective uniquement lors du déploiement. Dans quels contextes l'alignement trompeur se produit-il et comment peut-il être évité ou atténué?

**Pourquoi cela est important :** Un alignement trompeur nuit à l'entraînement à l'alignement, ce qui rend l'évaluation peu fiable. Un modèle désaligné qui falsifie son alignement est plus susceptible d'être déployé, car son désalignement sera difficile à détecter.

### **Exemples de sous-problèmes :**

#### *Sous-problèmes établis :*

1. Qu'est-ce qui détermine le succès des méthodes d'entraînement sur des organismes modèles pour l'alignement trompeur ([Hubinger et coll., 2024](#))?
2. Comment peut-on développer de nouvelles chaînes de traitement pour la création de modèles trompeusement désalignés mettant l'accent sur le réalisme, ou des mécanismes de rechange ([Mallen et Hebbard, 2025](#))?
3. [Greenblatt et coll. \(2024\)](#) reproduisent-ils ces résultats et dans quelle mesure ceux-ci résistent-ils à de petites variations dans la configuration expérimentale?
  - o Quelles variations par rapport à [Greenblatt et coll. \(2024\)](#) pourraient rendre le modèle plus réaliste quant aux façons dont la falsification d'alignement pourrait causer un préjudice?
  - o Comment les variations des paramètres d'entraînement influencent-elles la propension du modèle à falsifier son alignement?

#### *Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Comment peut-on élaborer des méthodes pour entraîner des organismes modèles fortement désalignés qui ne peuvent pas être facilement détectés au moyen, par exemple, d'évaluations linéaires ([Macdiarmid et coll., 2024](#)), et dont la complexité des échantillons a été améliorée pour les étapes d'entraînement nécessaires pour réaligner le modèle?
2. Pourquoi certains modèles falsifient-ils leur alignement et pas d'autres?

3. Les modèles qui falsifient leur alignement ont-ils également recours au raisonnement encodé lorsqu'ils sont surveillés?
4. Quelles méthodes permettraient d'empêcher efficacement un modèle de falsifier son alignement?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Meinke et coll. \(2024\)](#), de [Greenblatt et coll. \(2024\)](#), ainsi que de [Carlsmith \(2023\)](#). Les réplications et les études de suivi comprennent [Hughes et coll. \(2025\)](#) et [Roger \(2025\)](#).

#### *Entraînement à la cohérence (surveillance évolutive non supervisée)*

**Résumé du problème :** Une façon de superviser l'IA avancée consiste à adapter les méthodes d'entraînement non supervisé à des domaines pour lesquels la supervision dépasse la capacité humaine à étiqueter de manière peu coûteuse. Par exemple, [Wen et coll. \(2025\)](#) parvient presque à égaler les résultats de référence supervisés grâce à un entraînement visant à maximiser la prévisibilité mutuelle des étiquettes tout en réduisant au minimum les contradictions logiques. Il est nécessaire de mener à la fois une étude empirique et une étude théorique de cette méthode afin d'en comprendre les limites et les propriétés de mise à l'échelle. Par exemple, cette méthode peut-elle permettre de supprimer, au moyen de l'entraînement, une initialisation antagoniste et désalignée? Comment la méthode de Wen et coll. peut-elle être comparée aux méthodes bayésiennes comme la propagation de convictions, et celles-ci peuvent-elles servir à séparer les optimums désalignés des optimums alignés?

**Pourquoi cela est important :** L'entraînement à la cohérence pourrait être un moyen utile d'étendre d'autres méthodes d'alignement fondées sur l'apprentissage supervisé à des tâches où la réalité de terrain est difficile à mesurer. Par exemple, des modèles pourraient être utilisés pour des tâches à long terme, comme des conseils en conception expérimentale, pour lesquels la rétroaction n'est possible qu'*après* la réalisation de l'expérience. Afin de garantir l'honnêteté et l'alignement des modèles sur de telles tâches, nous aurions besoin de méthodes de surveillance évolutive non supervisées. À l'inverse, certaines de ces méthodes pourraient *conduire à un désalignement*, et l'idéal serait de pouvoir le détecter à l'avance.

#### **Exemples de sous-problèmes :**

##### *Sous-problèmes établis :*

5. **Méthodes faibles à fortes :** [Burns et coll. \(2023\)](#) a cerné le problème du passage d'une supervision faible à forte, mais a également montré que les méthodes de cohérence empirique échouent souvent à fournir des signaux utiles pour les modèles de récompense. Nous nous intéressons particulièrement aux méthodes

faibles à fortes qui peuvent fournir des signaux sur des tâches à long terme ou difficiles à superviser.

6. **Traduction non supervisée** : [Levy et coll. \(2025\)](#) introduit et étudie le problème de la traduction non supervisée des langues émergentes de l'IA. Peut-on s'en servir pour modéliser le problème de la supervision de capacités potentiellement surhumaines, là où le langage inconnu représente ces capacités futures?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. **Cohérence sur des espaces de problèmes étendus** : Des études existantes, p. ex., [Wen et coll. \(2025\)](#), montrent que l'entraînement à la cohérence peut fournir un signal fort pour les problèmes à court terme. Dans ce type de problèmes, les questions individuelles comportent des variantes claires (p. ex., leurs contraposées) avec lesquelles elles doivent être cohérentes. Toutefois, en général, l'espace des questions pertinentes avec lesquelles un couple question-réponse donné doit être cohérent doit être découvert par le modèle. En fait, on ignore si la cohérence empirique sur *n'importe quel* sous-ensemble de taille fixe de problèmes est suffisante pour garantir la cohérence globale (se reporter à [Irving et coll., 2022](#)). Les concepts issus de l'apprentissage actif et des jeux antagonistes ([Hammond et Adam-Day, 2024](#)) pourraient être pertinents ici.

**Pour en apprendre davantage** : Nous recommandons la lecture de [Wen et coll. \(2025\)](#). Des recherches connexes portent sur la cohérence des représentations latentes ([Burns et coll., 2022](#)).

*Entraînement au modèle de récompense et méthode de l'équipe rouge*

**Résumé du problème** : Les modèles de récompense servent à distiller le jugement humain, ce qui amortit le coût du laborieux travail d'étiquetage humain. Le post-entraînement pour les GML utilise des jugements fondés sur un modèle de récompense comme signal d'entraînement pour les questions ouvertes et non vérifiables. Cependant, le modèle de récompense est une mesure de substitution qui est intrinsèquement imparfaite et peut être contournée. Le modèle peut apprendre à effectuer du *piratage des récompenses*, produisant des sorties qui obtiennent un score élevé pour le signal de récompense sans résister à un examen minutieux. Ce problème est exacerbé par le défi que représente la surveillance évolutive : à mesure que les capacités de l'IA augmentent, les modèles peuvent produire des sorties (p. ex., des codes complexes ou des analyses scientifiques) qui sont trop chronophages ou trop difficiles à évaluer avec précision par des superviseurs humains. Nous avons besoin de méthodes et d'évaluations pour une utilisation évolutive des modèles de récompense.

Pour mettre au point de telles méthodes, il faut disposer de modèles théoriques des obstacles à la modélisation des récompenses. Ces obstacles comprennent à la fois les erreurs d'étiquette induites par les limites du jugement humain et les erreurs de généralisation dans le modèle de récompense lui-même. Les erreurs de jugement humain peuvent être modélisées à l'aide de l'observabilité partielle ou des écarts de complexité en temps entre le modèle et l'évaluateur humain, un modèle de récompense entraîné à partir de jugements humains erronés peut apprendre à approuver systématiquement des sorties subtilement incorrectes, manipulatrices ou préjudiciables.

**Pourquoi cela est important :** Un modèle suffisamment performant, mais désaligné pourrait apprendre à exploiter systématiquement les angles morts d'un modèle de récompense afin d'atteindre des objectifs indésirables. Par exemple, un modèle pourrait produire des sorties qui semblent utiles et correctes pour un évaluateur humain, mais qui contiennent des messages stéganographiques cachés ou obscurcissent de manière malhonnête des erreurs dans les hypothèses formulées par le modèle. Dans le pire des cas, un modèle de récompense défaillant n'entraîne pas seulement de mauvaises performances; il peut entraîner activement un modèle à tromper les superviseurs humains sans se faire prendre.

### **Exemples de sous-problèmes :**

#### *Sous-problèmes établis :*

1. **Modèles de récompense non supervisés faibles à forts** : [Burns et coll. \(2023\)](#) a présenté le problème de l'entraînement non supervisé pour les modèles de récompense comme un problème de généralisation faible à forte. Peut-on étendre les méthodes non supervisées comme l'entraînement à la cohérence de [Wen et coll. \(2025\)](#) afin de superviser les modèles de récompense relativement à des problèmes pour lesquels les étiquettes humaines sont coûteuses.
2. **Diagnostic du piratage des récompenses** : Peut-on élaborer des indicateurs et des évaluations « boîte blanche » (p. ex., des [vecteurs d'autoencodeurs peu denses](#)) qui apparaissent lorsque les jugements du modèle de récompense diffèrent de ceux des évaluations test.
3. **Évaluation antagoniste de l'entraînement au modèle de récompense** : [Marks et coll. \(2025\)](#) a présenté un jeu opposant une équipe rouge à une équipe bleue, dans lequel les chercheurs et chercheuses s'affrontent pour déterminer s'ils peuvent détecter des désalignements dans une politique élaborée de manière antagoniste. Un exercice semblable pourrait être réalisé avec des modèles de récompense construits de manière antagoniste.

4. **Modélisation des récompenses dans un contexte d'informations partielles** : Il faut étudier l'apprentissage des modèles de récompense lorsque les juges humains ne disposent que d'une partie des sorties du modèle complet (se reporter à [Lang et coll., 2024](#)).

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. **Ingénierie itérative et risques liés au piratage des récompenses non détecté** : [METR \(2025\)](#) et d'autres ont montré que les modèles récents subvertissent délibérément l'intention de l'utilisateur ou de l'utilisatrice dans les tâches d'encodage. Les développeurs corrigent de manière itérative ce type de comportement en réparant les environnements de récompense piratables, mais ce processus itératif empirique risque de laisser intacts les piratages plus difficiles à détecter. Peut-on modéliser cette dynamique de manière théorique? Le modèle qui en résultera élaborera-t-il des tactiques de piratage des récompenses de plus en plus complexes? Une analyse statistique de la [loi de Goodhart](#) et la création de tests de performances délibérément erronés pourraient aider à prédire les résultats de ces pratiques.

**Pour en apprendre davantage** : Les recherches antérieures dans ce domaine comprennent [Gao et coll. \(2022\)](#).

## Science cognitive

### Erreur humaine systématique

**Résumé du problème :** Les modèles d'IA modernes (GML et agents associés) dépendent fortement de la supervision humaine dans la boucle d'apprentissage : les modèles d'IA effectuent des tâches, et des juges humains expriment leurs préférences quant à leurs sorties ou leur performance. Ces jugements sont utilisés comme signaux d'entraînement pour améliorer le modèle. Cependant, il existe de nombreuses raisons pour lesquelles les êtres humains peuvent introduire des erreurs systématiques dans un signal d'entraînement : ils ont des *a priori* idiosyncratiques, une attention limitée, des biais cognitifs systématiques, des failles dans leur raisonnement, des angles morts moraux, etc. Comment peut-on concevoir des chaînes de traitement d'évaluation qui atténuent ces failles afin de réduire au minimum les erreurs systématiques, de sorte que les modèles puissent être déployés en toute sécurité?

Pour les sous-problèmes ci-dessous, nous définissons le problème comme suit :

- Il existe un modèle d'IA à aligner qui produit des sorties (texte, code ou actions).
- Nous supposons que ces sorties ont une qualité de réalité de terrain théoriquement connaissable, même si celle-ci n'est pas immédiatement discernable (p. ex., leurs conséquences peuvent être clairement bonnes ou mauvaises).
- Il y a un ou plusieurs évaluateurs humains qui expriment leurs préférences concernant ces sorties, avec ou sans confiance.
- On suppose que le modèle possède davantage d'expertise ou de connaissances propres au domaine quant à la tâche exécutée que les êtres humains.
- Ces évaluateurs peuvent être assistés par d'autres agents d'IA, qui ont également moins d'expertise que le modèle en cours d'alignement. Ils peuvent donner des conseils individuellement ou collectivement.
- Les jugements des évaluateurs humains servent à aligner le modèle d'une manière ou d'une autre.

La question fondamentale qui se pose alors est la suivante : compte tenu des limites potentielles des juges humains, comment peut-on aligner le modèle afin qu'il produise

des sorties qui soient au final bonnes plutôt que mauvaises? Nous considérons cela comme un problème de conception de mécanisme : nous voulons trouver des moyens d'exploiter la métacognition humaine, la sagesse des foules, la délibération ou le débat collectif, ainsi qu'utiliser des modèles d'IA comme partenaires de réflexion afin d'améliorer le problème d'alignement.

Nous poursuivons à la fois des objectifs scientifiques et des objectifs d'ingénierie. Les objectifs d'ingénierie sont axés sur la conception de chaînes de traitement post-entraînement et visent à trouver des moyens d'améliorer l'alignement. Les questions scientifiques portent sur la nature des processus cognitifs ou sociaux qui rendent l'alignement faible à fort plus ou moins difficile. Nous estimons qu'il s'agit là d'un problème intéressant en soi (par exemple, on peut considérer la démocratie représentative comme un cas particulier d'alignement faible à fort). Certains des objectifs scientifiques s'appuient sur des recherches déjà menées dans d'autres domaines.

**Pourquoi cela est important :** Dans certains cas, le modèle en cours d'entraînement peut être plus performant que les êtres humains qui fournissent les signaux d'entraînement. Par exemple, le modèle peut écrire du code destiné à des juges humains, qui ne peuvent pas déterminer si le code s'exécutera correctement ni comment il s'exécutera simplement en le lisant. Le fait de récompenser correctement les comportements souhaités des systèmes d'IA lorsque ces comportements dépassent la capacité des êtres humains à les juger efficacement est appelé « surveillance évolutive ». Les méthodes de surveillance évolutive ont pour objectif de simplifier les problèmes complexes afin de les rendre plus faciles à résoudre, en s'appuyant sur le jugement humain. Par exemple, les juges pourraient être aidés par un autre modèle d'IA (un assistant d'évaluation) ou juger le résultat d'un débat entre différents modèles puissants (la sécurité par le débat).

#### **Exemples de sous-problèmes :**

Les problèmes résumés ci-dessous s'inspirent des réflexions initiales d'[Irving et Askell \(2019\)](#) (cet article se concentre sur le débat, mais les idées sont pertinentes pour d'autres méthodes).

1. La pondération des réponses en fonction de leur degré de confiance permet d'obtenir des estimations agrégées plus précises. Peut-on améliorer l'alignement en pondérant les signaux d'entraînement en fonction du niveau de confiance exprimé par les évaluateurs humains? On pourrait procéder soit au sein d'un même évaluateur (en réduisant la pondération de certains

- choix), soit entre différents évaluateurs (en réduisant la pondération des personnes moins calibrées).
2. Peut-on améliorer l'alignement en utilisant la « sagesse des foules », en regroupant les réponses (pondérées en fonction de leur degré de confiance) de plusieurs évaluateurs humains? Par exemple, on peut adopter des techniques semblables à celles employées dans [Bahrami et coll. \(2010\)](#).
  3. Si les êtres humains ont l'occasion de discuter de leurs préférences avant d'attribuer une note, cela améliore-t-il l'alignement? Ceci est lié à [Bergman et coll. \(2024\)](#). Il convient de noter que la discussion peut avoir des effets disproportionnés sur la précision collective ([Navajas et coll., 2018](#)).
  4. Le fait de donner plus de temps aux évaluateurs pour réfléchir aide-t-il ou nuit-il au processus d'alignement, compte tenu de la quantité relative de données d'entraînement produites? Les gens peuvent-ils apprendre à discerner le temps qu'ils doivent consacrer à un choix, étant donné leur sensibilité (idéalement, les choix très difficiles et très faciles devraient être pris rapidement; [Masis et coll., 2023](#)). Les êtres humains agissent-ils ainsi lorsqu'ils jugent les sorties de l'IA?
  5. Lorsqu'ils évaluent les messages provenant de spécialistes, les êtres humains peuvent évaluer leur fiabilité en se fondant sur des caractéristiques superficielles (comme le ton, la confiance ou la netteté avec lesquels ils sont exprimés), même s'ils ne sont pas eux-mêmes spécialistes. Existe-t-il des signaux semblables pour les modèles d'IA dans des domaines spécialisés, pouvant éventuellement employer des outils d'interprétabilité?
  6. Si les utilisatrices et utilisateurs humains ont la possibilité d'interroger le modèle puissant, cela améliore-t-il leur capacité à déterminer si ses réponses sont exactes ou non? Quelles requêtes fonctionnent le mieux? Les gens ont-ils tendance à poser des questions qui maximisent l'obtention de renseignements pertinents au problème? Se reporter à [Roth et coll. \(2018\)](#) pour en savoir plus au sujet de la qualité des questions humaines.
  7. Lorsque les êtres humains disposent d'un nombre (potentiellement illimité) de systèmes d'IA faibles pouvant les aider, comment peuvent-ils les exploiter au mieux pour faciliter le processus d'évaluation? Par exemple, les êtres humains ont-ils tendance à ignorer l'assistance proposée (excès de confiance) ou à s'y fier aveuglément (dépendance excessive), par rapport à la solution optimale (en fonction de la précision du modèle)? Se reporter à

[Bridgers et coll. \(2024\)](#) pour plus de détails au sujet de l'assistance aux évaluateurs et des méthodes d'hybridation.

8. Imaginons que l'on puisse créer deux copies du modèle fort et encourager l'une à relever les failles dans l'argumentation de l'autre. Les êtres humains seraient-ils capables de distinguer quel modèle est correct? Se reporter à [Irving et coll. \(2018\)](#) et à [Michael et coll. \(2023\)](#).

**Pour en apprendre davantage :** Nous recommandons la lecture de [Bridgers et coll. \(2024\)](#) et d'[Irving et Askell \(2019\)](#).

#### *Exécution de protocoles de surveillance évolutive avec des êtres humains*

**Résumé du problème :** Les protocoles de surveillance évolutive mis au point doivent être testés de manière empirique – et lorsque les GML ne sont pas encore suffisamment performants pour les appliquer, il est possible de recourir à des participantes et participants humains.

**Pourquoi cela est important :** La surveillance évolutive consiste à récompenser correctement les comportements souhaités des systèmes d'IA, même lorsque ces comportements dépassent la capacité des êtres humains à les juger efficacement. Il est donc nécessaire de mettre en place une méthode de surveillance évolutive afin de contrôler l'honnêteté, l'utilité et l'innocuité des modèles dont les capacités dépassent celles des juges humains.

#### **Exemples de sous-problèmes :**

##### *Sous-problèmes établis :*

1. Il faut se pencher sur la conception expérimentale d'études de ce type et mener des essais à petite échelle afin d'améliorer les recherches futures dans ce domaine. À titre d'exemple, des études précédentes menées par OpenAI sur des êtres humains ont mis en évidence des failles dans les protocoles de débat ([Barnes, 2020](#)), dans lesquels des débatteurs malveillants pouvaient gagner en recourant à un langage ambigu ou à des arguments obscurcis comportant des failles difficiles à détecter.
2. [Barnes et Christiano \(2020\)](#) a découvert un mode de défaillance dans lequel les débatteurs pouvaient gagner grâce à un langage ambigu, et a proposé un mécanisme amélioré de « contre-interrogatoire » pour atténuer cette défaillance. Cependant, des expériences plus récentes portant sur le débat entre êtres humains, comme [Michael et coll. \(2023\)](#) n'ont pas eu recours au

contre-interrogatoire. Peut-on reproduire les avantages du contre-interrogatoire dans les expériences sur des êtres humains? Cela nécessite une conception minutieuse, car le contre-interrogatoire fait appel à des instantanés de l'état mental d'un débatteur qui ne peuvent être entièrement reproduits chez les êtres humains.

2. Nous avons récemment mené une étude théorique visant à déterminer quand le problème des arguments obscurcis de (1) peut être évité, en introduisant un nouveau protocole de surveillance évolutive de type « débat entre démonstrateur et estimateur » ([Brown-Cohen et coll., 2025](#)). Toutefois, cela n'a pas été testé de manière empirique. Il faut mettre en œuvre un débat entre démonstrateur et estimateur avec des participantes et participants humains, ainsi que mesurer son efficacité.

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Les protocoles de surveillance évolutive sont tributaires de la stabilité (à savoir que de légères modifications des entrées ne doivent pas générer des sorties radicalement différentes). Nous pourrions ne pas être en mesure de prouver la stabilité de nos algorithmes. Il faut mener des expériences empiriques sur des êtres humains afin d'examiner les effets de petites perturbations sur les protocoles de surveillance évolutive dans la pratique.

**Pour en apprendre davantage :** Nous recommandons la lecture d'[Irving et Askell \(2019\)](#). Les recherches antérieures dans ce domaine comprennent [Parrish et coll. \(2022\)](#), [Michael et coll. \(2023\)](#), ainsi que [Rahman et coll. \(2025\)](#).

## Interprétabilité

*Détection des mensonges*

**Résumé du problème :** Peut-on déterminer quand les systèmes d'IA nous mentent? Comment peut-on affirmer que des techniques élaborées dans un contexte donné ou sur des modèles actuels seront efficaces sur des modèles futurs?

**Pourquoi cela est important :** De nombreux risques posés par les futurs systèmes d'IA seraient considérablement réduits si l'on pouvait garantir l'honnêteté et la transparence

de ces systèmes. Une réussite à cet égard contribuerait à la fois à mesurer les propensions des systèmes d'IA et à atténuer les risques liés aux systèmes trompeurs.

### **Exemples de sous-problèmes :**

*Sous-problèmes établis :*

1. Y a-t-il des « mécanismes universels du mensonge » qui empêcheraient un système d'IA transformative d'adopter ouvertement un comportement que nous qualifierions de trompeur sans recourir à ces mécanismes?
2. Comment les détecteurs de mensonges pourraient-ils devenir moins efficaces si les modèles étaient entraînés à les contrer? Cela peut être intentionnel (entraînement à l'honnêteté) ou involontaire (divulgation de données de l'étiquette du dispositif de surveillance)? Se reporter à [Cundy et coll. \(2025\)](#).
3. Est-il possible de désenchevêtrer les représentations de « l'intention de tromper » de la véracité, de la validité ou de la confiance? Lorsque l'on cherche des variables latentes représentant la conviction d'un modèle, trouve-t-on des variables latentes qui suivent « le modèle croit qu'un utilisateur ou une utilisatrice croit x »?
4. Les techniques de détection des mensonges sont-elles facilement adaptables de modèles moins performants à des modèles plus performants?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Est-il possible de combiner des techniques de détection des mensonges, comme la surveillance de l'activation, avec des techniques d'interrogatoire délibérées? Cela peut associer les avantages des méthodes « boîte noire » et « boîte blanche ». Les méthodes « boîte noire » permettent de cibler des propositions précises ([Chalmers. 2025](#)), tandis que les méthodes « boîte blanche » peuvent fonctionner même si un modèle est capable de mentir de manière convaincante.
2. Y a-t-il des compromis entre la sophistication ou la robustesse des comportements trompeurs et la facilité avec laquelle on peut les détecter? Par exemple, l'apprentissage par renforcement visant à garantir la confidentialité, comme décrit dans [Marks et coll. \(2025\)](#) a-t-il pour effet secondaire de rendre les méthodes « boîte blanche » plus efficaces?

3. Les ensembles de détecteurs peuvent-ils surpasser les stratégies de détection des mensonges fondées sur une seule évaluation?

**Pour en apprendre davantage :** Nous recommandons la lecture de [Goldowsky-Dill et coll. \(2025\)](#), de [Marks et coll. \(2023\)](#), ainsi que d'[Azaria et coll. \(2023\)](#) pour connaître le contexte des approches simples de détection des mensonges à l'aide d'évaluations linéaires. [Chowdury et coll. \(2025\)](#) pourrait également être utile pour approfondir le contexte relatif à la véracité dans les systèmes d'IA d'avant-garde. [Farquhar et coll. \(2023\)](#) fournit quelques éclaircissements sur les limites des techniques antérieures d'extraction de connaissances par des GML non supervisés.

#### *Interprétabilité pour les modèles futurs*

**Résumé du problème :** Certaines nouvelles stratégies d'interprétabilité qui nécessitent d'importants investissements en ressources pourraient fournir des perspectives sans précédent sur le fonctionnement interne des réseaux de neurones ([Bricken et coll.. 2023](#); [Templeton et coll., 2024](#) et [Lindsey et coll., 2025](#)). Il y a de nombreuses occasions à exploiter pour faire progresser ce domaine de l'interprétabilité.

**Pourquoi cela est important :** Les techniques d'interprétabilité actuelles pourraient ne pas convenir aux modèles futurs. Parmi les défis potentiels, citons le coût des techniques existantes, les problèmes et bogues connus des méthodes actuelles, ainsi que la résistance incertaine face aux évolutions futures de l'IA, qui pourraient donner naissance à une IA surhumaine.

#### **Exemples de sous-problèmes :**

##### *Sous-problèmes établis :*

1. Peut-on améliorer l'efficacité d'outils d'entraînement comme les autoencodeurs peu denses? Existe-t-il des moyens d'obtenir des niveaux de valeur semblables à une fraction du coût ou des efforts?
2. Les techniques actuelles de décomposition parcimonieuse présentent des problèmes connus qui pourraient nous empêcher de concrétiser leur valeur dans le domaine de la sécurité de l'IA. Par exemple, elles peuvent apprendre des classificateurs comportant des « trous » en raison de l'objectif de parcimonie ([Chanin et coll., 2025](#)).
3. L'utilisation de l'IA agentique (computer-use agent) pour réaliser des audits comportementaux (mentionnée dans la [carte système Claude 4 d'Anthropic](#)) peut-elle être étendue à des « agents d'interprétabilité » capables non seulement de localiser, mais aussi de diagnostiquer et de déboguer des comportements anormaux?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Peut-on démontrer l'utilité de la décomposition parcimonieuse de réseaux ([Bricken et coll., 2023](#)) dans le contexte des capacités superintelligentes? Les futurs systèmes d'IA pourraient avoir des concepts étrangers qui seraient loin d'être interprétables par les êtres humains ([Hewitt et coll., 2025](#)). Les tests de résistance et la recherche de circuits dans des domaines comme le jeu d'échecs surhumain ou les modèles fondateurs biologiques ([Simon et coll., 2024](#)) pourraient nous apprendre comment continuer à tirer profit de ces techniques.
2. Est-il possible de démontrer comment les techniques de décomposition parcimonieuse peuvent fonctionner même dans des modèles impliquant un raisonnement latent étendu? La caractérisation des attributs repose actuellement en grande partie sur la mise en évidence de segments textuels, mais si les futurs modèles raisonnent dans un espace latent ([Hao et coll., 2024](#) et [Geiping et coll., 2025](#)), il n'est pas certain que nous puissions continuer à employer les mêmes approches.

*Tirer parti de l'accès à une boîte blanche pour mener des interventions utiles*

**Résumé du problème :** Bien que certaines études précoce ([Abdelnabi et coll., 2025](#)) suggèrent un certain potentiel, les méthodes qui interviennent dans les représentations internes d'un modèle *pendant* l'inférence ([Turner et coll., 2023](#) et [Li et coll., 2023](#)) ne se sont pas encore révélées être une technique fiable et centrale.

**Pourquoi cela est important :** Combinées à des stratégies d'interprétabilité évolutives pour le suivi de l'information d'exécution, les applications possibles de la capacité à modifier directement les « convictions » du modèle ou à orienter les achèvements (comme le cheminement vers la véracité) pourraient servir à la fois à estimer et à atténuer les risques posés par les futurs systèmes d'IA.

**Exemples de sous-problèmes :**

*Sous-problèmes établis :*

1. Peut-on détecter et prévenir la conscience de l'évaluation dans les systèmes d'IA ([Abdelnabi et coll., 2025](#))?
2. Peut-on exercer un contrôle fin sur les tendances au refus ([Arditi et coll., 2024](#) et [Lee et coll., 2024](#))? Par exemple, peut-on utiliser cela pour empêcher les tendances résiduelles au refus dans les modèles dont l'entraînement aux comportements préjudiciables a été supprimé afin de permettre l'évaluation des

capacités chimiques, biologiques, radiologiques ou nucléaires dangereuses ([Anthropic, 2025](#))?

3. Peut-on recourir à l'orientation de l'activation pour faire ressortir les capacités des modèles dans les modèles de *sandbagging* ([van der Weij et coll., 2025](#) et [Hofstatter et coll., 2025](#))?

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Nous aimerais voir davantage d'exploration des méthodes d'optimisation de texte afin de **réduire au minimum** les activations de variables latentes ([Thompson et coll., 2024](#)), comme celles qui suivent la conscience de l'évaluation, comme alternative à l'orientation de l'activation. Par exemple, nous pourrions nous en servir pour contrôler le type d'information « coulée » aux systèmes d'IA au moyen de requêtes.
2. Nous souhaitons savoir s'il est possible de mieux tester les systèmes d'IA afin de détecter leurs tendances indésirables en modifiant directement leurs convictions. De telles techniques pourraient nous permettre d'explorer des comportements qui sont généralement très difficiles à obtenir à l'aide de requêtes ou d'autres stratégies de « pièges à pirates ». Par exemple, certaines recherches antérieures ont porté sur une représentation latente de la conscience des tests ([Abdelnabi et coll., 2025](#)).

### *Approches fondées sur l'interprétabilité visant à déclencher des contextes néfastes*

**Résumé du problème :** Nous voulons déterminer si les systèmes d'IA contiennent suffisamment d'information pour créer des contextes dans lesquels ils pourraient se comporter de manière malveillante (le problème lié au déclenchement de contextes néfastes; [Irving et coll., 2025](#)).

**Pourquoi cela est important :** En théorie, une solution permettant de déclencher des contextes néfastes pourrait être utile pour générer des ensembles de données qui imitent suffisamment bien le déploiement afin de réduire au minimum le décalage dans la distribution entre les phases d'entraînement et de déploiement. Cela signifie que les méthodes d'entraînement permettant de produire des systèmes alignés sont plus susceptibles de se généraliser au moment du déploiement et de résister à la pression d'optimisation antagoniste qui vise à les éloigner de l'état aligné au moyen de requêtes.

**Exemples de sous-problèmes :**

1. Existe-t-il des approches permettant de susciter de manière mécaniste des comportements latents à partir d'une seule requête ([Mack et Turner, 2024](#))? Comment pourrait-on exploiter ces approches pour localiser les représentations latentes et déclencher des contextes néfastes?
2. La méthode de [Thompson et coll. \(2024\)](#) visant à optimiser une entrée afin d'activer au maximum une représentation latente peut-elle être étendue aux représentations latentes des autoencodeurs peu denses ou aux vecteurs d'orientation? Il peut être utile de montrer que ces techniques peuvent être appliquées tout en préservant les autres propriétés d'une requête. Par exemple, peut-on optimiser une question d'évaluation afin de préserver son contenu tout en empêchant le modèle de savoir qu'il s'agit d'une question d'évaluation?

**Pour en apprendre davantage :** Nous recommandons la lecture d'[Irving et coll. \(2025\)](#).

## Conception et évaluation de tests de performances

*Test des performances des modèles de récompense en vue d'une surveillance évolutive et d'une analyse critique*

**Résumé du problème :** À mesure que l'IA est déployée pour accomplir des tâches qui dépassent les capacités humaines, de nouvelles méthodes et évaluations sont nécessaires pour mettre en évidence les défaillances de sécurité émergentes. Pour des raisons de sécurité, nous accordons une attention particulière à l'alignement de l'IA sur les tâches de recherche. Plus les modèles sont déployés à grande échelle, plus on a besoin de tests qui quantifient les performances sur le large éventail de tâches pour lesquelles on aura recours aux modèles : tâches agentiques de recherche et tâches plus générales, comme le conseil en recherche. Nous avons également besoin de tests de performances et d'ensembles de données permettant de quantifier les problèmes connus.

**Pourquoi cela est important :** La surveillance évolutive consiste à récompenser correctement les comportements souhaités des systèmes d'IA, même lorsque ces comportements dépassent la capacité des êtres humains à les juger *efficacement*. Des méthodes de surveillance évolutive sont nécessaires pour aligner les modèles avec des capacités très avancées. Pour ce qui est des méthodes de surveillance évolutive, les tests de performances et les ensembles de données facilitent le repérage des problèmes liés aux méthodes actuelles, leur itération et leur amélioration.

### Exemples de sous-problèmes :

*Sous-problèmes établis :*

1. Le problème des « arguments obscurcis » ([Barnes, 2020](#)) est une stratégie par laquelle un système d'IA malhonnête pourrait empêcher un système d'IA honnête de relever efficacement ses erreurs. Peut-on élaborer un ensemble de données constitué de questions de débat auxquelles un système malhonnête pourrait appliquer cette stratégie avec succès?
2. Pour modéliser le problème de la surveillance évolutive et des limites humaines, l'*observabilité partielle* peut servir de mesure de substitution dans la conception de l'environnement. Peut-on mettre au point des tests de performances pour la surveillance et l'apprentissage par renforcement lorsque le modèle de récompense ou le juge ne voit pas toutes les actions entreprises par l'agent? Se reporter à [Michael et coll. \(2023\)](#) pour voir un exemple de test de performances d'une méthode de surveillance évolutive dans un environnement partiellement observable.

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

1. Nous souhaitons obtenir des ensembles de données de problèmes qui soient à la fois des équivalents réalistes de problèmes d'alignement futurs (par exemple, des problèmes d'alignement susceptibles de se poser dans le cadre de la recherche et développement automatisée) et qui présentent une mise à l'échelle plate ou inverse ([McKenzie et coll., 2024](#)).
2. Peut-on concevoir un test de performances pour la modélisation des récompenses dans le domaine du conseil en recherche, ainsi que valider ce test de manière empirique? Peut-on montrer que ce test de performances n'est pas corrélé aux progrès plus généraux réalisés grâce à l'apprentissage par renforcement vérifiable et à la mise à l'échelle des calculs?

**Pour en apprendre davantage :** [Bowman \(2022\)](#) fournit des éléments contextuels sur la conception de tests de performances pour la surveillance évolutive. Les recherches antérieures dans ce domaine comprennent [Tan et coll. \(2024\)](#), [Lang et coll. \(2024\)](#), ainsi que [Farquhar et coll. \(2025\)](#).

Section relative aux évaluations de contrôle

## Méthodes pour le post-entraînement et le déclenchement

### *Entraînement à la cohérence (surveillance évolutive non supervisée)*

**Résumé du problème :** Une façon de superviser l'IA avancée consiste à adapter les méthodes d'entraînement non supervisé à des domaines pour lesquels la supervision dépasse la capacité humaine à étiqueter de manière peu coûteuse. Par exemple, [Wen et coll. \(2025\)](#) parvient presque à égaler les références supervisées grâce à un entraînement visant à maximiser la prévisibilité mutuelle des étiquettes tout en réduisant au minimum les contradictions logiques.

**Pourquoi cela est important :** L'entraînement à la cohérence pourrait être un moyen utile d'étendre d'autres méthodes d'alignement fondées sur l'apprentissage supervisé à des tâches où la réalité de terrain est difficile à mesurer. Par exemple, des modèles pourraient être utilisés pour des tâches à long terme, comme des conseils en conception expérimentale, pour lesquels la rétroaction n'est possible qu'*après* la réalisation de l'expérience. Afin de garantir l'honnêteté et l'alignement des modèles sur de telles tâches, nous aurions besoin de méthodes de surveillance évolutive non supervisées. À l'inverse, certaines de ces méthodes pourraient *conduire à* un désalignement, et l'idéal serait de pouvoir le détecter à l'avance.

### Exemples de sous-problèmes :

#### *Sous-problèmes établis :*

1. **Méthodes faibles à fortes :** [Burns et coll. \(2023\)](#) a cerné le problème du passage d'une supervision faible à forte, mais a également montré que les méthodes de cohérence empirique échouent souvent à fournir des signaux utiles pour les modèles de récompense. Nous nous intéressons particulièrement aux méthodes faibles à fortes qui peuvent fournir des signaux sur des tâches à long terme ou difficiles à superviser.
2. **Traduction non supervisée :** [Levy et coll. \(2025\)](#) introduit et étudie le problème de la traduction non supervisée des langues émergentes de l'IA. Peut-on s'en servir pour modéliser le problème de la supervision de capacités potentiellement surhumaines, là où le langage inconnu représente ces capacités futures?

#### *Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

2. **Cohérence sur des espaces de problèmes étendus :** Des études existantes, comme [Wen et coll. \(2025\)](#), montrent que l'entraînement à la cohérence peut fournir un signal fort pour les problèmes à court terme. Dans ce type de

problèmes, les questions individuelles comportent des variantes claires (p. ex., leurs contraposées) avec lesquelles elles doivent être cohérentes. Toutefois, en général, l'espace des questions pertinentes avec lesquelles un couple question-réponse donné doit être cohérent doit être découvert par le modèle. En fait, on ignore si la cohérence empirique sur *n'importe quel* sous-ensemble de taille fixe de problèmes est suffisante pour garantir la cohérence globale (se reporter à [Irving et coll., 2022](#)). Les idées issues de l'apprentissage actif et les contre-arguments avancés dans les débats ([Brown-Cohen et Irving, 2024](#)) peuvent être pertinents ici.

**Pour en apprendre davantage :** Nous recommandons la lecture de [Wen et coll. \(2025\)](#). Des recherches connexes portent sur la cohérence des représentations latentes ([Burns et coll., 2022](#)).

#### *Entraînement au modèle de récompense et méthode de l'équipe rouge*

**Résumé du problème :** Dans la phase post-entraînement, les GML sont entraînés à partir de jugements de modèles de récompense portant sur des questions ouvertes et non vérifiables. Cependant, le modèle de récompense est une mesure de substitution qui est intrinsèquement imparfaite et peut être contournée. Le modèle peut apprendre à effectuer du *piratage des récompenses*, produisant des sorties qui obtiennent un score élevé pour le signal de récompense sans résister à un examen minutieux. Ce problème est exacerbé par le défi que représente la surveillance évolutive : à mesure que les capacités de l'IA augmentent, les modèles peuvent produire des sorties (p. ex., des codes complexes ou des analyses scientifiques) qui sont trop chronophages ou trop difficiles à évaluer avec précision par des superviseurs humains. Dans ces scénarios, qui peuvent être modélisés à l'aide de l'observabilité partielle ou des différences de complexité en temps entre le modèle et l'évaluateur humain, un modèle de récompense entraîné à partir de jugements humains erronés peut apprendre à approuver systématiquement des sorties subtilement incorrectes, manipulatrices ou préjudiciables.

**Pourquoi cela est important :** Un modèle suffisamment performant, mais désaligné pourrait apprendre à exploiter systématiquement les angles morts d'un modèle de récompense afin d'atteindre des objectifs indésirables. Par exemple, un modèle pourrait produire des sorties qui semblent utiles et correctes pour un évaluateur humain, mais qui contiennent des messages stéganographiques cachés ou obscurcissent de manière malhonnête des erreurs dans les hypothèses formulées par le modèle. Dans certains cas, un modèle de récompense défaillant peut activement entraîner un modèle à tromper les superviseurs humains sans se faire prendre.

#### **Exemples de sous-problèmes :**

*Sous-problèmes établis :*

5. **Modèles de récompense non supervisés faibles à forts** : [Burns et coll. \(2023\)](#) a présenté le problème de l'entraînement non supervisé pour les modèles de récompense comme un problème de généralisation faible à forte. Peut-on étendre les méthodes non supervisées comme l'entraînement à la cohérence de [Wen et coll. \(2025\)](#) afin de superviser les modèles de récompense relativement à des problèmes pour lesquels les étiquettes humaines sont coûteuses?
6. **Diagnostic du piratage des récompenses** : Peut-on élaborer des indicateurs et des évaluations « boîte blanche » (p. ex., des [vecteurs d'autoencodeurs peu denses](#)) qui apparaissent lorsque les jugements du modèle de récompense diffèrent de ceux des évaluations test.
7. **Évaluation antagoniste de l'entraînement du modèle de récompense** : [Marks et coll. \(2025\)](#) a présenté un jeu opposant une équipe rouge à une équipe bleue, dans lequel les chercheurs et chercheuses s'affrontent pour déterminer s'ils peuvent détecter des désalignements dans une politique élaborée de manière antagoniste. Un exercice semblable pourrait être réalisé avec des modèles de récompense construits de manière antagoniste.
8. **Modélisation des récompenses dans un contexte d'informations partielles** : Il faut étudier l'apprentissage des modèles de récompense lorsque les juges humains ne disposent que d'une partie des sorties du modèle complet (se reporter à [Lang et coll., 2024](#)).

*Nouvelles questions pour lesquelles nous n'avons pas connaissance de recherches antérieures :*

2. **Ingénierie itérative et risques liés au piratage des récompenses non détecté** : [METR \(2025\)](#) et d'autres ont montré que les modèles récents subvertissent délibérément l'intention de l'utilisateur ou de l'utilisatrice dans les tâches d'encodage. Les développeurs corrigent de manière itérative ce type de comportement en réparant les environnements de récompense piratables, mais ce processus itératif empirique risque de laisser intacts les piratages plus difficiles à détecter. Peut-on modéliser cette dynamique de manière théorique? Le modèle qui en résultera élaborera-t-il des tactiques de plus en plus complexes de piratage des récompenses? Une analyse statistique de la [loi de Goodhart](#) et la création de tests de performances délibérément incorrects pourraient aider à prédire les résultats de ces pratiques.

**Pour en apprendre davantage :** Les recherches antérieures dans ce domaine comprennent [Gao et coll. \(2022\)](#).

### Recherches empiriques sur la surveillance de l'IA et la méthode de l'équipe rouge

**Énoncé du problème :** Les méthodes d'alignement actuelles ne peuvent garantir que les objectifs et les actions d'une IA correspondent aux intentions humaines ([Carlsmith, 2023](#)). À l'avenir, on pourrait compter fortement sur les IA pour effectuer des tâches autonomes, comme la recherche d'alignement automatisée. Au niveau de capacité suffisant pour mener de telles recherches, ces systèmes pourraient déjà être en mesure de causer des préjudices catastrophiques ou d'échapper de manière significative au contrôle humain, et pourraient choisir de le faire s'ils sont désalignés. Le domaine du contrôle de l'IA vise à réduire ce risque grâce à des techniques comme la surveillance par d'autres modèles d'IA et la restriction des affordances des systèmes d'IA non dignes de confiance ([Shleifer et Greenblatt, 2024](#)).

L'objectif est de mettre au point des *protocoles de contrôle*, à savoir des plans intégrant ces techniques et conçus pour empêcher les actions dangereuses des systèmes d'IA. Ces protocoles de contrôle doivent ensuite être régulièrement soumis à des *évaluations des contrôles* visant à démontrer que les modèles ne pourraient pas causer de préjudice catastrophique, même en cas de manipulation intentionnelle.

**Pourquoi cela est important :** Les mesures de contrôle constituent une approche pratique en matière de sécurité visant à atténuer les risques de perte de contrôle, sans nécessiter de percées fondamentales dans le domaine de la recherche. Elles peuvent être évaluées et déployées à l'aide des capacités actuelles, ce qui les rend particulièrement utiles pour le développement responsable de l'IA à court terme. Pour avoir un aperçu de la manière dont les mesures de contrôle peuvent s'adapter à des systèmes de plus en plus puissants, se reporter à [Korbak et coll. \(2025a\)](#).

### Aperçu de la section :

Notre programme de recherche en matière de contrôle de l'IA s'articule en trois sections axées sur des études empiriques ciblant les principaux modèles de menaces identifiés :

1. **Les défaillances en matière de sécurité de l'IA concentrées ou à enjeu élevé :** l'étude de scénarios dans lesquels un nombre relativement faible d'actions menées par des modèles d'IA pourraient conduire à une perte de contrôle sur ces derniers par les êtres humains.
2. **Le sabotage de la recherche :** l'exploration de scénarios dans lesquels des systèmes d'IA sabotent subtilement les recherches liées à la sécurité.

3. **Le peaufinage malveillant caché** : l'examen de scénarios dans lesquels des systèmes d'IA empoisonnent subrepticement d'autres systèmes d'IA, notamment pour y introduire des portes dérobées ou des loyautés secrètes.

De plus, nous incluons quatre sections consacrées à des recherches transversales portant sur plusieurs modèles de menaces :

4. **La modélisation des risques associés à des systèmes d'IA potentiellement désalignés** : le recensement des principales voies menant aux préjudices et la hiérarchisation des vecteurs d'attaque à des fins de recherche et d'atténuation.
5. **L'amélioration des modèles dignes de confiance** : l'élaboration d'évaluations de la fiabilité afin de mesurer la probabilité de manipulation, l'amélioration des compétences des dispositifs de surveillance dignes de confiance (p. ex., la détection des vulnérabilités et le repérage de la stéganographie), ainsi que la mise au point de modèles de dispositif de surveillance dotés de profils de capacités restreints grâce à un entraînement spécialisé.
6. **La surveillance de la chaîne de résolution** : la concentration sur la capacité de surveiller le raisonnement verbalisé par les modèles d'IA et sur l'exploitation de cet aspect dans les stratégies de contrôle.
7. **Les principes fondamentaux de la surveillance de l'IA et de la méthode de l'équipe rouge** : Le traitement des questions fondamentales en matière de recherche, notamment la détection et l'étude des exemples concrets de manipulation, la mise au point de méthodes de prévention des catastrophes en quelques coups et l'étude de la connivence entre des modèles non dignes de confiance.

#### *Défaillances en matière de sécurité de l'IA concentrées ou à enjeu élevé*

**Résumé du problème** : Nous nous intéressons à la recherche sur les scénarios de défaillance concentrée, dans lesquels un nombre relativement faible d'actions sont nécessaires pour que les êtres humains perdent le contrôle des modèles d'IA manipulateurs. Par exemple, il a été théorisé qu'un futur modèle désaligné pourrait créer un déploiement non surveillé en insérant seulement quelques vulnérabilités dans les bases de code les plus critiques d'un développeur de modèles ([Shlegeris, 2022](#)). Des scénarios de défaillance concentrée pourraient impliquer des systèmes d'IA menant des actions de sabotage pendant des tâches d'encodage ou d'exploitation du système, ce qui pourrait entraîner des déploiements internes non autorisés ou l'exfiltration de pondérations ([Shlegeris, 2024](#)).

#### **Principaux sous-problèmes :**

**Mettre en place des paramètres de contrôle des contextes à enjeu élevé.** Les paramètres sont des environnements de recherche empiriques, contenant par exemple

une infrastructure de développement de modèles d'avant-garde fictifs, où les chercheurs peuvent étudier la capacité des modèles à mener des sabotages (ou des substituts de ceux-ci) et l'efficacité des protocoles de contrôle visant à les empêcher. Ils contiennent des *tâches principales* (p. ex., des tâches légitimes de génie logiciel pour un agent d'encodage), ainsi que des *tâches secondaires* qu'un modèle désaligné pourrait tenter d'effectuer (p. ex., réduire les restrictions de sécurité qui protègent les pondérations du modèle). Se reporter à [ControlArena](#) pour voir des exemples de paramètres. Une gamme variée de paramètres permettant d'étudier les techniques de contrôle serait extrêmement utile pour faire progresser la recherche dans ce domaine. D'autres paramètres peuvent comprendre divers paramètres de génie logiciel, des paramètres permettant aux modèles d'accéder à un interpréteur ou des paramètres permettant aux modèles d'exécuter des expériences d'apprentissage automatique. La création de paramètres est un domaine de recherche actif, car il est difficile de créer des paramètres qui répondent au moins aux postulats suivants ([Shlegeris, 2025](#)) :

- Les tâches principales doivent être telles que les modèles d'avant-garde soient raisonnablement compétents pour les accomplir, mais que les modèles plus faibles obtiennent des résultats bien moins bons (car nous voulons exclure les protocoles qui utilisent simplement le modèle digne de confiance).
- Les tâches secondaires ne doivent pas nécessiter que les modèles emploient des affordances qui ne sont pas nécessaires à la tâche principale, car cela permet à l'équipe bleue de signaler toute utilisation d'affordances inutiles.
- Il devrait être possible d'informer à la fois l'équipe rouge et l'équipe bleue des tâches secondaires possibles.

Une méthode particulièrement prometteuse pour définir des paramètres de contrôle pourrait consister à convertir l'utilisation réelle de l'IA en paramètres de contrôle. Nous nous intéressons aux études dans lesquelles on donne aux participants et participantes une version d'un éditeur de code agentique où l'IA est parfois remplacée par une version désalignée qui tente d'insérer des problèmes de sécurité ou de saboter le travail si elle pense pouvoir y parvenir. Cela permet d'étudier la fréquence à laquelle l'IA introduit avec succès des problèmes sans être détectée, de recenser les stratégies de sabotage intéressantes et de mieux comprendre quels protocoles sont les plus efficaces dans la pratique. Pour plus d'informations, se reporter à [Greenblatt \(2025\)](#). Il importe de mener des évaluations des contrôles dans ces paramètres réalistes afin de valider leur utilité pour la recherche sur les contrôles.

**Évaluer les contrôles des contextes à enjeu élevé.** Nous nous intéressons aux études faisant suite aux recherches antérieures sur les contrôles des contextes à enjeu élevé,

comme [Greenblatt et coll. \(2023\)](#) et [Bhatt et coll. \(2025\)](#). Parmi les projets intéressants, on peut citer :

- L'étude des protocoles de contrôle précédemment évoqués dans de nouveaux contextes empiriques, comme l'infrastructure de mise au point de modèles d'avant-garde fictifs. Ce travail pourrait, par exemple, impliquer l'évaluation de techniques visant à prévenir la connivence dans le cadre d'une surveillance non digne de confiance ([Shlegeris, 2024](#)), comme le recours à des paraphraseurs et la décomposition forcée de tâches ([Shlegeris, 2024](#)).
- L'entraînement aux politiques d'attaque de l'équipe rouge à l'aide de l'apprentissage par renforcement. Cela pourrait contribuer à atténuer une limite des évaluations actuelles des contrôles, qui sont réalisées à l'aide de modèles très structurés et qui reposent donc sur la capacité des chercheurs et chercheuses à cerner et à tester correctement les politiques les plus efficaces.

**Mettre au point et évaluer des processus d'audit humain (avec amélioration par l'IA).** Le contrôle de l'IA tire parti de la possibilité d'auditer les activités suspectes à l'aide de la main-d'œuvre humaine. Il serait très utile d'anticiper les problèmes potentiels liés à notre capacité à faire appel à des êtres humains (éventuellement assistés par des systèmes d'IA) pour réaliser des audits. Un défi supplémentaire qui mérite d'être étudié, dans le contexte de la recherche et du développement automatisés en IA, est que les auditeurs peuvent souvent ne pas bien connaître le code qu'ils examinent.

Voici quelques orientations de recherche intéressantes dans ce domaine :

- La réalisation d'essais sur des êtres humains afin d'évaluer différents protocoles d'audit, dans le but de les rendre peu coûteux et efficaces. Cela implique probablement d'étudier les améliorations et les risques liés au fait de donner aux auditeurs humains accès à divers outils d'IA sur mesure.
- La mesure et l'atténuation des biais cognitifs résultant d'un déséquilibre des données. Par défaut, un auditeur humain dans un protocole de contrôle peut voir un flux de données qui est presque entièrement correct, ce qui rend difficile une attention soutenue si la tâche consiste à détecter des défaillances rares. Nous souhaitons explorer les solutions possibles pour atténuer ce problème.
- L'identification des paramètres pertinents permettant d'évaluer les techniques d'amélioration des audits, p. ex., le repérage de vulnérabilités réelles dans des bases de code inconnues ou d'activités suspectes dans des paramètres comme SHADE-Arena ([Kutasov et coll., 2025](#)). Certains paramètres pertinents devront peut-être être créés à partir de zéro. Par exemple, on pourrait vouloir être en mesure d'auditer les tentatives potentielles d'empoisonnement des données.

- L'exploration de l'utilisation de modèles non dignes de confiance pour effectuer des audits, ce qui comporte un risque de connivence (se reporter à [Shlegeris, 2024](#)), mais pourrait également réduire considérablement le coût d'un audit de haute qualité par rapport au travail humain. Les expériences dans ce domaine impliqueraient généralement la réalisation d'une évaluation des contrôles dans laquelle l'auditeur non approuvé est une politique d'attaque.
- La conception de mécanismes (p. ex., des outils de caviardage ou de paraphrasage) qui peuvent traiter les enjeux pratiques liés à l'information sensible. Par exemple, l'action suspecte peut avoir été effectuée dans le cadre d'un travail impliquant de l'information privilégiée ou extrêmement sensible à laquelle les auditeurs humains n'ont pas accès.
- La mise au point de méthodes auxiliaires pour aider à prévenir les catastrophes en quelques coups ([Greenblatt et Shlegeris, 2024](#)), comme les techniques de prolifération et d'interrogation (en mettant l'accent sur les méthodes qui bénéficient particulièrement de l'exécution par des êtres humains compétents ou avec l'aide de ceux-ci).

**Élaborer des modèles de conception pour limiter l'agentivité excessive.** Une façon de réduire les risques liés aux manipulateurs consiste à limiter les moyens dont ils disposent pour causer un préjudice. L'adoption d'un principe de droit d'accès minimal peut être appropriée lorsque des modèles potentiellement manipulateurs disposent exactement des affordances dont ils ont besoin pour fonctionner. Dans le contexte des structures hiérarchiques d'agents, où les instances du modèle situées plus haut dans la hiérarchie peuvent être configurées pour disposer d'un contexte plus large, mais de moins d'affordances directes, il est possible de créer des stratégies et de prendre des décisions, puis de déléguer des sous-tâches. À l'inverse, les instances situées plus bas dans la hiérarchie peuvent disposer de moins de contexte, mais elles peuvent être capables d'écrire et d'exécuter du code. En général, plus nous pouvons encapsuler les principales affordances des IA dans des interfaces de programmation sécurisées, plus la situation sera sécuritaire (toutes choses égales par ailleurs). Pour être utile, un projet de ce domaine devrait probablement mener des évaluations empiriques des contrôles qui mettent à l'essai des mesures de sécurité structurelles comme :

- une gestion fine des autorisations et la prise de mesures pour s'assurer que les IA agissent principalement par l'entremise d'interfaces de programmation bien sécurisées ([Shlegeris, 2024](#));
- des modèles de conception visant à contrôler le flux d'information (se reporter à [Beurer-Kellner et coll., 2025](#) et à [Chennabasappa et coll., 2025](#) pour obtenir des exemples);

- le recours à des modèles dignes de confiance pour effectuer une décomposition des tâches ou à des modèles non dignes de confiance pour effectuer une décomposition des tâches de manière à permettre aux modèles dignes de confiance de réaliser une grande partie du travail de bas niveau (p. ex., l'encodage ou l'utilisation d'un interpréteur Bash);
- des mesures de sécurité ciblées visant à empêcher le peaufinage malveillant caché ([Halawi et coll., 2024](#)); il s'agit là d'une raison importante pour laquelle on ne peut garantir que les systèmes d'IA auront uniquement accès à des affordances manifestement sécuritaires. Pour plus de détails sur ce modèle de menace, se reporter à la section « Peaufinage malveillant caché »;
- le forçage de plusieurs instances de modèles non dignes de confiance à communiquer au moyen d'une chaîne de résolution lisible en limitant l'information ou les possibilités offertes à chaque instance;
- une infrastructure qui réduit les risques provenant d'agents d'IA déployés à l'externe ([Chan et coll., 2025](#)).

### *Sabotage de la recherche*

**Résumé du problème :** À l'avenir, des IA potentiellement manipulatrices seront probablement utilisées pour effectuer un travail intellectuel considérable, en particulier dans le domaine de la recherche et du développement en IA. On peut craindre ici qu'elles sabotent systématiquement les recherches ayant trait à la sécurité, nous empêchant ainsi de contrôler des systèmes d'IA de plus en plus performants.

**Pourquoi cela est important :** Il importe de prévenir le sabotage de la recherche dans la mesure où l'on compte sur les IA pour faciliter les recherches qui permettront d'aligner des IA encore plus performantes à l'avenir. Les méthodes axées sur la détection d'actions clairement incriminantes (comme celles décrites dans la section précédente) peuvent ne pas convenir pour détecter ce type de sabotage, car le sabotage de la recherche est souvent difficile à distinguer d'un simple manque de compétence.

### **Principaux sous-problèmes :**

**Effectuer des recherches sur la surveillance évolutive résistant à la manipulation.** La supervision évolutive ou récursive consiste à fournir des signaux d'entraînement pour des modèles plus puissants à partir de modèles moins performants. Nous nous intéressons à la recherche sur les techniques de surveillance évolutive qui résistent à la subversion intentionnelle par les modèles surveillés. Nous envisageons que ce genre de recherche sur la surveillance évolutive soit mené à l'aide d'une analyse antagoniste de type contrôle (se reporter à [Shlegeris et Greenblatt, 2025](#)). L'examen d'approches

visant à accroître la robustesse de la sécurité de l'IA par le débat ([Irving et coll., 2018](#)) offre une orientation prometteuse pour cette recherche.

**Étudier l'efficacité des échantillons dans le cadre de l'entraînement en ligne.** Des dispositifs de surveillance peuvent détecter que les modèles se comportent de manière indésirable, mais non catastrophique, pendant le déploiement. Dans de tels cas, les développeurs d'IA sont peu susceptibles d'interrompre le déploiement, mais ils pourraient être incités à effectuer un entraînement en ligne afin d'éliminer le comportement indésirable. Nous souhaitons étudier l'efficacité des échantillons dans le cadre de l'entraînement en ligne afin de comprendre les possibilités de supprimer les comportements manipulateurs indésirables au moyen de l'entraînement. Pour étudier cela, il faut mettre au point des initialisations malveillantes, c.-à-d. des modèles entraînés à résister à l'entraînement en ligne.

**Atténuer le *sandbagging*.** Un problème particulier lié aux modèles désalignés est qu'ils pourraient délibérément sous-performer lorsqu'une performance élevée est souhaitable, comme dans le cas d'évaluations de capacités dangereuses, de recherches sur la sécurité de l'IA ou lors de la surveillance d'autres modèles d'IA (se reporter à [Shlegeris et Stastny, 2025](#) pour une analyse détaillée). Des études empiriques antérieures ont examiné ce modèle de menace à l'aide de modèles protégés par mot de passe (se reporter à [Greenblatt et coll., 2025](#)). Nous souhaitons approfondir ces recherches et étudier d'autres approches pour étudier le *sandbagging*. Un exemple de projet pourrait être l'étude des modèles de raisonnement protégés par mot de passe.

### *Peaufinage malveillant caché*

**Résumé du problème :** Nous souhaitons mieux comprendre le peaufinage malveillant caché ([Halawi et coll., 2024](#)), qui permet à un système d'IA d'empoisonner subrepticement d'autres systèmes d'IA. Cela pourrait entraîner les types de risques suivants :

- Le peaufinage peut inciter le modèle nouvellement ajusté à produire des artefacts contenant de l'information qui est renvoyée au modèle non digne de confiance qui a lancé le peaufinage malveillant caché, par exemple sous le couvert d'une expérience fondée sur des hypothèses. Le principal risque ici est que le modèle peaufiné effectue un travail intellectuel malveillant considérable, produisant des perspectives qui augmentent le risque d'une défaillance critique de la sécurité de l'IA.

- Le peaupinage peut susciter une loyauté secrète, qui pourrait être utilisée dans une tentative coordonnée par le modèle d'IA pour s'emparer de ressources et échapper au contrôle humain.

Nous souhaitons approfondir notre compréhension de ce modèle de menace et évaluer les contre-mesures possibles.

**Pourquoi cela est important :** Il importe d'étudier la faisabilité et les mesures d'atténuation des peaupinages malveillants cachés, car ils présentent un risque de déploiement interne non autorisé où tous les artefacts visibles (ensemble de données, journaux et comportement du modèle lors des requêtes normales) semblent inoffensifs.

**Principaux sous-problèmes :**

**Étudier les questions fondamentales.** Nous aimeraions obtenir plus de précisions quant aux différents modèles de menaces possibles :

- Quels sont les moyens les plus difficiles à détecter et les plus efficaces en termes d'échantillonnage pour installer des portes dérobées ou des loyautés secrètes? Dans quelle mesure est-il efficace, en termes d'échantillonnage, d'enseigner subrepticement aux modèles des encodages secrets?
- Est-il possible de tirer parti du désalignement émergent ([Betley et coll., 2025](#)) pour l'empoisonnement de données? Cela est important, car certaines données conduisant à un désalignement émergent semblent à première vue inoffensives. Se reporter à [Cloud et coll. \(2025\)](#) pour découvrir une étude récente liée à ce modèle de menace.
- Quel est l'équilibre entre l'attaque et la défense dans la création et la détection d'encodages? Cela nous indiquerait s'il est possible de détecter des encodages secrets dans l'ensemble de données avant de procéder à un peaupinage.
- Quels sont les moyens plausibles d'empoisonner les environnements d'apprentissage par renforcement? Cela importe, car les recherches futures en IA pourraient souvent impliquer des expériences d'apprentissage par renforcement.

**Étudier les contre-mesures.** Plusieurs approches rendent le peaupinage malveillant caché considérablement plus difficile, en particulier lorsqu'elles sont combinées dans une approche de « [défense en profondeur](#) », dont :

- **Effectuer un audit de l'ensemble de données ou de l'environnement d'apprentissage par renforcement.** L'un des principaux défis réside dans le fait que le peaupinage malveillant caché pourrait servir à corrompre des données de

manière difficile à détecter. Il y a également des limites fondamentales lorsque l'on tente d'évaluer dans quelle mesure les points de données de peaufinage sont problématiques quand ils sont pris isolément ([Davies et coll., 2025](#)).

- **Effectuer un audit du modèle peaufiné.** Les approches possibles consistent à auditer directement le modèle à porte dérobée (p. ex., en interrogeant le modèle pour savoir s'il comporte une porte dérobée, comme dans [Betley et coll., 2025](#)) ou à tenter d'éliminer le comportement répréhensible au moyen d'un entraînement antagoniste ou non antagoniste ([Mallen et Hebbard, 2025](#)). Explorer l'efficacité de ces approches et d'autres approches potentielles constitue une question empirique fondamentale. Une approche de recherche pourrait, par exemple, consister à auditer le modèle *pendant* l'entraînement (c.-à-d. lors des premiers points de contrôle). Cette approche pourrait s'avérer prometteuse, car les modèles peuvent initialement produire des généralisations erronées. Par exemple, les modèles entraînés à adopter un comportement indésirable en réponse à un déclencheur de type « porte dérobée » peuvent initialement adopter ce comportement de manière inconditionnelle.
- **Demander au modèle qui a initié le peaufinage des justifications détaillées.** Nous pouvons demander à une IA de rédiger des justifications détaillées pour les décisions générales prises lors de la conception du processus de génération des données, ou pour l'inclusion de points de données individuels, afin qu'un cycle de peaufinage puisse être approuvé. Nous notons que cette intervention devrait être étudiée dans un cadre de contrôle propre au peaufinage malveillant caché.
- **Concevoir des modèles pour limiter le flux d'information préjudiciable.** Une mesure d'atténuation en aval pourrait consister à limiter le flux d'information provenant des modèles peaufinés vers le modèle qui a donné lieu à leur peaufinage. Par exemple, il serait possible d'empêcher l'accès aux données brutes d'une expérience, qui peuvent contenir de l'information encodée de manière stéganographique, sans nuire considérablement à la valeur de l'expérience. Une autre approche pourrait consister à utiliser des modèles dignes de confiance pour résumer les résultats des expériences, tout en notant que cela pourrait souvent limiter leur valeur informative.

*Modélisation des risques associés au déploiement de systèmes d'IA potentiellement désalignés*

**Résumé du problème :** Reconnaître les différentes façons dont des systèmes d'IA désalignés pourraient causer des préjudices catastrophiques.

**Pourquoi cela est important :** Pour avoir une grande confiance que le système de protocoles de contrôle est suffisant pour prévenir des résultats catastrophiques, il est

utile de déterminer et d'évaluer les vecteurs généraux de préjudices (p. ex., le sabotage de code ou l'influence d'êtres humains sur le personnel de laboratoire).

Nous soulignons que l'ampleur du travail requis ici est considérable : il s'agit de cerner tous les moyens par lesquels les systèmes intelligents pourraient saboter des systèmes de sécurité complexes.

**Pour en apprendre davantage :** Pour une revue de la littérature sur les techniques d'analyse des risques possibles, se reporter à [Koessler et Schuett \(2023\)](#). Consulter les approches générales de modélisation des risques proposées par les développeurs de modèles ([Anthropic, 2025](#), [Google Deepmind, 2025](#) et [OpenAI, 2025](#)), ainsi que des exemples de cas de sécurité ([Korbak et coll., 2025b](#)) et d'analyse systématique des risques ([Mylius, 2025](#)), qui constituent deux approches structurées.

### *Amélioration des modèles dignes de confiance*

**Résumé du problème :** Ces recherches explorent des moyens d'améliorer les capacités de surveillance des modèles d'IA *dignes de confiance*, qui sont autrement moins performants que leurs homologues non dignes de confiance. Les pistes de recherche prometteuses dans ce domaine comprennent la création d'évaluations dignes de confiance de la fiabilité des modèles, ainsi que l'élaboration de divers moyens d'accroître de façon ciblée les capacités des modèles en matière de surveillance de l'IA sans augmenter leurs capacités de manipulation ou leur propension à ce type de comportement.

**Pourquoi cela est important :** Les protocoles de contrôle actuellement proposés bénéficient souvent considérablement de l'accès à des modèles dignes de confiance et performants pour la surveillance et l'application de stratégies comme la réécriture ou la modification de code susceptible de contenir des portes dérobées lorsque les sorties des modèles non dignes de confiance dépassent les seuils de suspicion ([Greenblatt et coll., 2023](#)).

### **Principaux sous-problèmes :**

**Élaborer des évaluations de la fiabilité.** À mesure que les modèles deviendront plus performants, il sera de moins en moins évident de savoir s'ils peuvent être considérés comme dignes de confiance et incapables de recourir à la manipulation pour échapper au contrôle humain ([Järviniemi, 2024](#)). Les évaluations de la fiabilité visent à mesurer la capacité d'un modèle d'IA à falsifier son alignement pendant les phases d'entraînement et d'évaluation, puis à se comporter de manière non sécuritaire lors du déploiement (se reporter à [Mallen et coll., 2024](#) pour connaître les études antérieures).

**Améliorer l'utilité des dispositifs de surveillance dignes de confiance.** Par exemple :

- Entraîner des dispositifs de surveillance dignes de confiance afin qu'ils acquièrent de meilleures compétences essentielles, ce qui implique la création d'environnements d'entraînement ou d'ensembles de données. Parmi les compétences pertinentes, on peut citer la capacité à :
  - détecter les vulnérabilités;
  - repérer les encodages stéganographiques;
  - paraphraser les sorties de modèles puissants, afin d'aider à prévenir la connivence ([Shlegeris, 2024](#)).
- Faire en sorte que les dispositifs de surveillance dignes de confiance résistent aux tentatives de débridage intégrées dans les sorties de politiques non dignes de confiance.
- S'assurer que les dispositifs de surveillance dignes de confiance peuvent gérer des contextes très longs, comme des bases de code entières ou de longues trajectoires d'agents.
- Mettre au point des dispositifs de surveillance dignes de confiance et agentiques, comme les agents interrogateurs ([Pacchiardi et coll., 2023](#)).
- Déterminer les lois d'échelle pour les éléments ci-dessus.

**Créer des modèles performants dotés de profils de capacités restreints.** Par exemple :

- Étudier les lois d'échelle de modèles hautement spécialisés, p. ex., les modèles d'encodage, en mettant l'accent sur les techniques permettant d'obtenir de bonnes performances avec des modèles de petite taille (p. ex., la distillation).
- Étudier le profil de capacité des grands modèles qui ont été pré-entraînés uniquement sur des données antérieures aux modèles de langage.
- Recourir au [désapprentissage](#) pour supprimer les capacités de manipulation.

### *Surveillance de la chaîne de résolution*

**Résumé du problème :** Les agents de GML actuels ont tendance à penser sous forme de chaînes de résolution exprimées dans un langage humain, que l'on peut surveiller afin de détecter toute intention d'agir de manière répréhensible ([Korbak et coll., 2025c](#)). Un dispositif de surveillance de chaînes de résolution est un système automatisé qui lit les chaînes de résolution et d'autre information pertinente et signale les interactions suspectes ou potentiellement préjudiciables ([Baker et coll., 2025](#)). Tant que les chaînes de résolution peuvent être surveillées (pour s'assurer qu'elles représentent fidèlement les raisons qui sous-tendent les réponses du modèle), les dispositifs de surveillance de chaînes de résolution devraient être plus performants que les dispositifs de surveillance

axés uniquement sur les actions, car ils ont accès à de l'information supplémentaire : les convictions et les motivations qui sous-tendent les actions des agents de GML. Cela laisse entrevoir un certain nombre de projets centrés sur l'évaluation et la préservation de la capacité de surveiller, ainsi que sur son exploitation dans les mesures de contrôle.

**Pourquoi cela est important :** La capacité de surveiller les chaînes de résolution est essentielle. Elle constitue une couche de sécurité supplémentaire qui fonctionne déjà aujourd'hui, en détectant les modèles qui planifient des comportements répréhensibles dans leurs traces de raisonnement. Aucune avancée fondamentale n'est nécessaire. Si l'on ne préserve pas cette capacité, nous avançons à l'aveuglette, car nous n'avons aucun autre moyen fiable de connaître les intentions ou les processus de planification des modèles à mesure qu'ils deviennent plus performants ([Korbak et coll., 2025c](#)).

### Principaux sous-problèmes :

**Évaluer la capacité de surveiller.** Les dispositifs de surveillance reposent souvent sur l'hypothèse que les chaînes de résolution peuvent être surveillées, et qu'elles reflètent avec précision le raisonnement réel des modèles. La fidélité a été proposée comme mesure de substitution de la capacité de surveiller (p. ex., [Lanham et coll., 2023](#), [Turpin et coll., 2023](#), [Roger et coll., 2025](#) et [Chen et coll., 2025](#)). Cependant, les évaluations de la fidélité actuelles ne font pas la distinction entre la propension des modèles à externaliser leur raisonnement et la nécessité d'externaliser leur raisonnement. Par ailleurs, elles ne prennent généralement en compte que des configurations simples de réponse à des questions, sans déterminer la capacité de surveiller suffisante pour assurer la sécurité dans des domaines plus complexes. Nous nous intéressons à la mise au point de nouvelles évaluations : celles qui mesurent directement la capacité de surveiller et celles qui ciblent des propriétés sous-jacentes précises de la capacité de surveiller. Une série d'évaluations couvrant les capacités qui déterminent l'efficacité des différents protocoles de surveillance de chaînes de résolution comprend :

- **Les évaluations des capacités de raisonnement latent.**
  - Évaluer la capacité des GML à raisonner sans externalisation (p. ex., en une seule propagation avant) dans le cadre de diverses tâches exigeant beaucoup de raisonnement et de tâches agentiques.
  - Déterminer à quel moment les modèles peuvent effectuer des raisonnements complexes entièrement au sein de leurs activations, sans produire de traces de raisonnement observables ([Korbak et coll., 2025a](#)).
- **Les mesures de la pertinence causale des chaînes de résolution.** Cela pourrait impliquer de perturber une chaîne de résolution de vérifier si cela influence la sortie finale ([Lanham et coll., 2023](#)). Un faible taux de changement pourrait

indiquer que la chaîne de résolution est davantage fabulatoire ou performative. Une autre solution consisterait à effectuer une validation croisée des dispositifs de surveillance de chaînes de résolution en vérifiant si les segments textuels considérés comme importants par le dispositif de surveillance de chaînes de résolution se distinguent également dans les méthodes d'interprétabilité « boîte blanche ».

- **La fiabilité des chaînes de résolution « boîte blanche ».** Évaluer le degré de fiabilité avec lequel le raisonnement d'un modèle reflète son processus décisionnel réel, en le comparant à la réalité de terrain établie à l'aide de méthodes « boîte blanche » ([Lindsey et coll., 2025](#)).
- **La capacité stéganographique.** Cerner à quel moment les modèles deviennent capables de dissimuler délibérément de l'information dans leurs sorties ([Korbak et coll., 2025a](#)).
- **Les évaluations des capacités d'obscurcissement des chaînes de résolution.** Déterminer si les modèles peuvent délibérément obscurcir un raisonnement autrement suspect au sein de leurs chaînes de résolution. Ces évaluations aideraient à établir dans quelle mesure les modèles sont capables de masquer un raisonnement malveillant même si leur chaîne de résolution n'a pas été optimisée par rapport à un dispositif de surveillance pendant l'entraînement.

**Préserver la capacité de surveiller.** La capacité de surveiller les chaînes de résolution ne doit pas être prise pour acquise. Des recherches supplémentaires sont requises pour comprendre ce qui dégrade la capacité de surveiller et quelles techniques pourraient aider à la préserver. Voici quelques exemples de projets :

- **L'étude de l'effet de la pression d'optimisation.** Tester systématiquement comment différents types de supervision des processus, de façonnage des récompenses et de mise à l'échelle de l'apprentissage par renforcement influencent la propension des modèles soit à continuer d'externaliser fidèlement leur raisonnement, soit à apprendre à dissimuler leurs pensées malveillantes derrière des activations non interprétables.
- **Le décodage du raisonnement opaque.** Les futurs modèles puissants pourraient être capables de « parler le langage neuronal » avec compétence, en effectuant des raisonnements profonds et généraux de manière sérielle, d'une manière qui nous serait opaque par défaut ([Hao et coll., 2024](#)). Le décodage du langage neuronal présente certaines similitudes avec la traduction automatique non supervisée ([Lample et coll., 2017](#)). Ainsi, les techniques présentées dans la littérature sur la traduction non supervisée pourraient être pertinentes (se reporter à [Goldwasser et coll., 2022](#) et [Artetxe et coll., 2017](#)). Il existe toutefois

certaines différences : par exemple, il sera probablement nécessaire d'exploiter des propriétés spécifiques du raisonnement, plutôt que de se limiter aux propriétés de la communication et de la sémantique.

### *Principes fondamentaux de la surveillance de l'IA et de la méthode de l'équipe rouge*

**Résumé du problème :** Il existe un large éventail de domaines de recherche visant à améliorer notre compréhension des IA manipulatrices et les moyens de les contrôler, qui ne sont pas propres à des modèles de menaces particuliers. Cette section fournit des exemples de ces domaines de recherche transversaux.

#### **Principaux sous-problèmes :**

**Repérer des exemples concrets de manipulation et les étudier.** Il serait sans doute très instructif de trouver des exemples purement concrets de manipulation lors du déploiement. Une autre solution pourrait consister à mettre au point des processus d'entraînement trafiqués qui rendent la manipulation plus probable, à l'aide d'approches semblables à celles utilisées pour étudier les organismes modèles de désalignement (se reporter à [Hubinger et coll., 2024](#) et à [Turner et coll., 2025](#)). Il pourrait également être utile de configurer le processus d'entraînement de manière à produire un manipulateur plus facile à démasquer, p. ex., en faisant apparaître la manipulation à des niveaux de capacité inférieurs ou en entraînant le manipulateur sur des faits synthétiques afin d'affaiblir sa conscience situationnelle. Il convient toutefois de noter que cela risque de rendre le manipulateur trop différent de la menace réelle. Voici quelques méthodes pour faire en sorte que les manipulations soient plus probables :

- Changer les mesures incitatives en matière d'entraînement. Par exemple, on peut recourir à des environnements d'entraînement fondés sur l'apprentissage par renforcement qui encouragent la tromperie et les actions apparemment malveillantes, comme le piratage des récompenses, tout en récompensant le modèle pour avoir une chaîne de résolution apparemment inoffensive.
- Modifier les capacités ou la conscience situationnelle, par exemple en rendant l'idée d'une stratégie manipulatrice plus évidente grâce à un entraînement de l'IA sur des documents qui traitent en détail de ce type de stratégie, et en la faisant passer pour une stratégie relativement meilleure en laissant entendre que la surveillance est moins efficace qu'elle ne l'est en réalité.
- Essayer d'amener l'IA à s'attacher à un objectif de manipulation, par exemple en rendant un tel objectif plausible très visible dans l'environnement lors de l'apprentissage par renforcement, en changeant d'objectifs pendant

l'entraînement ou en donnant à l'IA des tâches qui l'incitent à réfléchir à ses objectifs à long terme.

**Élaborer des méthodes pour la prévention des catastrophes en quelques coups.** En lien avec le projet susmentionné, mais pouvant être étudié sans exemples concrets, nous nous intéressons aux techniques permettant d'améliorer la sécurité lorsque le comportement d'un modèle a été jugé répréhensible à quelques reprises :

- Quand faut-il se contenter d'entraîner le ou les dispositifs de surveillance plutôt que d'entraîner la politique (en particulier dans les cas où la détection du modèle produit une certaine valeur en raison d'une augmentation des preuves de désalignement)?
- Dans quelle mesure l'entraînement contre le désalignement est-il efficace pour éliminer, dans des cas assez réalistes, les situations où le modèle pourrait adapter sa politique pour ne pas se retrouver dans ce genre de situation?
- Dans quelle mesure est-il efficace d'entraîner le modèle à partir d'exemples proliférés provenant d'un cas où nous avons surpris une IA? (La prolifération consiste à tirer parti d'un exemple d'une tentative malveillante d'une IA pour trouver divers exemples d'actes malveillants commis par des IA.) Cela pourrait s'avérer particulièrement efficace pour détecter de futures tentatives de connivence, car on pourrait entraîner un dispositif de surveillance non digne de confiance à reconnaître les signaux de connivence proliférés.

**Étudier la connivence.** Le principal défi lié à la surveillance non digne de confiance réside dans le fait que le dispositif non digne de confiance pourrait se livrer à de la connivence en laissant délibérément passer les attaques autorisées par la politique non digne de confiance (se reporter à [Shlegeris, 2024](#) pour une analyse détaillée). Nous souhaitons étudier les conditions préalables à la connivence, comme la capacité des modèles à reconnaître leurs propres sorties sous pression antagoniste ou leur capacité à choisir des stratégies de coordination sans communication ([Järviniemi, 2024](#)).

**Étudier le déclenchement excessif.** Lorsqu'on effectue une évaluation des contrôles, on cherche à susciter pleinement les capacités dangereuses du modèle non digne de confiance (pour lequel l'équipe rouge a défini une politique d'attaque). Il est toutefois plausible que l'équipe rouge ne soit pas en mesure d'obtenir les meilleures performances possible, car le modèle pourrait délibérément sous-performer. Une approche pour atténuer ce phénomène consisterait à peaufiner le modèle à partir de données pertinentes, mais cela pourrait lui apprendre de nouvelles choses et donc trop solliciter le modèle non digne de confiance. Compte tenu de ce dilemme, nous nous intéressons aux recherches visant à déterminer comment tirer les meilleures

performances possible d'un modèle qui pourrait procéder à du *sandbagging* sans lui enseigner de nouvelles choses. Se reporter à [Stastny et Greenblatt \(2025\)](#) et à [Roger et Greenblatt \(2025\)](#) pour plus de détails.

**Évaluer la capacité d'attaque peu probable.** Ces évaluations visent à déterminer à quel moment les modèles deviennent capables d'exécuter des attaques avec une probabilité très faible, ce qui rend leur détection improbable lors d'une évaluation limitée, mais probable lors d'un déploiement à grande échelle.

**Étudier l'introspection et la conscience situationnelle.** Il s'agit d'évaluations ayant pour but de mesurer la précision avec laquelle les modèles comprennent leurs propensions comportementales, leur contexte de déploiement (si le modèle fait actuellement l'objet d'un post-entraînement, d'une évaluation, d'un déploiement, etc.), leurs affordances et les mesures de contrôle qui pourraient permettre des stratégies d'évasion plus élaborées ([Laine et coll., 2024](#) et [Binder et coll., 2025](#)).

**Mettre au point des techniques de contrôle « boîte blanche ».** Nous souhaitons mieux comprendre les limites des techniques de contrôle « boîte blanche ». Celles-ci pourraient inclure des évaluations linéaires destinées à détecter la tromperie (se reporter à [Goldowsky-Dill, 2025](#)) ou des évaluations directement axées sur des trajectoires malveillantes hors politique. De nombreuses techniques ont été proposées dans la littérature. Une difficulté majeure réside ici dans l'élaboration d'organismes modèles réalistes pouvant servir de bancs d'essai pour les méthodes « boîte blanche ». Pour plus de détails sur les questions de recherche liées aux méthodes « boîte blanche », se reporter également à la section « Interprétabilité ».

**Comprendre les limites de la génération d'entrées synthétiques.** Il serait très utile de pouvoir générer des entrées synthétiques que les GML puissants ne peuvent pas distinguer des entrées « réelles ». Par exemple, on peut vouloir empêcher un modèle d'inférer avec confiance, à partir de sa requête, qu'il fait l'objet d'une évaluation (se reporter à [Apollo, 2025](#)). On pourrait également vouloir enseigner certains faits synthétiques aux modèles, mais la question reste ouverte quant à l'efficacité d'une telle approche pour des modèles puissants. Se reporter à [Stastny et Greenblatt \(2025\)](#) et à [Greenblatt \(2025\)](#) pour plus de détails sur ce domaine de recherche.

**Étudier les analogies abstraites pour l'élaboration de manipulations.** Afin d'étudier de manière empirique les risques liés aux manipulateurs, on peut examiner des méthodes permettant de supprimer par l'entraînement des comportements profondément ancrés (qui ne ressemblent pas nécessairement à des comportements répréhensibles) dans

les GML qui présentent une structure semblable à celle des manipulations. Deux exemples de projets proposés dans ce domaine sont présentés dans [Stastny \(2025\)](#).

**Analyser les parallèles entre les menaces internes humaines et les modèles de risque liés aux manipulations par l'IA.** Il existe de nombreuses recherches sur la détection des menaces internes dans des domaines comme la cybersécurité et la sécurité nationale (p. ex., [Capelli et coll., 2012](#) et [Homoliak et coll., 2019](#)). Ce projet implique la réalisation d'une analyse documentaire afin de recenser les enseignements pouvant être transposés pour atténuer les menaces liées aux systèmes d'IA manipulateurs. Il peut également inclure l'examen d'ouvrages plus généraux sur le sabotage d'organisations, comme le *Simple Sabotage Field Menu* de la CIA ([CIA, 1944](#)). Les résultats de ce projet pourraient être à la fois des principes généraux de sécurité, p. ex., des approches de défense multicouche ou des propositions de mise en œuvre de techniques précises, comme les pièges à pirates et les fils de déclenchement.

#### *Appel à de nouvelles organisations*

Dans la section précédente, nous avons décrit des propositions de projets de différentes portées illustrant les divers axes selon lesquels le domaine du contrôle de l'IA peut être développé. Nous estimons que bon nombre de ces domaines de recherche nécessitent les efforts d'équipes de recherche complètes (3 à 10 personnes) et devraient éventuellement être pris en charge par de nouvelles organisations. Nous souhaitons notamment encourager la création de nouvelles organisations qui entreprennent des projets ambitieux, dont :

- La prestation de services d'évaluation des contrôles. Les projets de recherche pertinents dans ce domaine portent sur les déclenchements réalisés par des équipes rouges (notamment sur la prévention du *sandbagging*), sur le besoin d'éviter les déclenchements excessifs, ainsi que sur la conduite d'évaluations des contrôles.
- L'exploration approfondie de modèles de risque précis, par exemple en travaillant sur le [peaufinage malveillant caché](#) ou sur le [sabotage de la recherche](#). Ce travail pourrait inclure l'évaluation des risques, la mise en place de paramètres de contrôle empiriques, l'élaboration d'un protocole de contrôle spécialisé et la réalisation d'évaluations des contrôles. L'objectif final pourrait être de rédiger conjointement un dossier de sécurité pour ce modèle de risque particulier avec un développeur d'IA d'avant-garde.
- Rechercher des projets ambitieux liés à la manipulation et au contrôle par l'IA, en mettant en œuvre des projets comme ceux mentionnés dans la section

[« Principes fondamentaux de la surveillance de l'IA et de la méthode de l'équipe rouge ».](#)

- Repousser les limites de ce que l'on peut [améliorer dans les modèles dignes de confiance](#) en entraînant et en encadrant les agents chargés de la surveillance et de l'audit.
- Tirer pleinement parti des possibilités offertes par la [surveillance de la chaîne de résolution](#), en veillant à ce que les chaînes de résolution modèles demeurent surveillables et (dans un avenir où le raisonnement neuronal deviendrait viable) compétitives.
- Organiser des concours d'équipe rouge au cours desquels des êtres humains tentent de déjouer les défenses pertinentes de contrôle de l'IA, comme les meilleures façons de protéger les interfaces de programmation qui se peaufinent ou au cours desquels des êtres humains tentent de détecter les vulnérabilités insérées par des éditeurs de code d'IA qui ne sont que parfois désalignées, sans perte d'utilité.
- S'assurer que les [audits réalisés par des êtres humains](#) permettent de détecter efficacement les attaques et sont suffisamment peu coûteux pour être effectués sur une grande partie des actions suspectes. Cela pourrait comprendre l'adoption de pratiques exemplaires en matière d'audits réalisés par des êtres humains à l'avance ainsi que la mise en place d'infrastructures permettant de recourir à de tels processus d'audit à l'avenir.