Exemple de Cadre de Sécurité et de Sûreté (Projet)

Ce document est un exemple de Cadre de Sécurité et de Sûreté, rédigé du point de vue d'un développeur hypothétique d'IA de pointe visant à traiter les risques catastrophiques potentiels qui pourraient résulter du développement et du déploiement de modèles avancés. Il s'inspire principalement de la section Sécurité et Sûreté du Troisième Projet de Code de Pratique et secondairement des "Éléments communs des politiques de sécurité de l'IA de pointe", plutôt que de représenter la vision de METR d'un cadre de sécurité idéal. Le Cadre n'a pas été examiné pour sa conformité juridique avec le Code de Pratique. Ce document a été traduit de l'anglais à l'aide d'un grand modèle de langage, il peut donc contenir des erreurs de traduction.

Formats: en (PDF, DOCX), fr (PDF, DOCX), zh-Hans (PDF, DOCX)

Résumé

Ce Cadre décrit nos engagements pour évaluer et atténuer les risques systémiques qui peuvent découler des modèles d'IA de pointe que nous développons. Il fournit une méthodologie pour déterminer quand les risques atteignent des niveaux inacceptables, mettre en œuvre des mesures techniques appropriées de sécurité et de sûreté, et assurer une gouvernance adéquate tout au long du cycle de vie du modèle d'IA.

Cadre d'évaluation des risques systémiques

Le Cadre identifie quatre catégories de risques systémiques, que nous évaluons dans nos modèles et nous engageons à atténuer.

- **Capacités de cyberoffensive** : Permettre des cyberattaques à un niveau de sophistication exigeant typiquement des experts humains bien dotés en ressources

- **Risque chimique, biologique, radiologique et nucléaire (CBRN)** : Conseils de niveau expert qui pourraient permettre le développement et la prolifération d'armes CBRN dangereuses
- **Perte de contrôle** : Comportements du modèle qui pourraient contourner la supervision humaine ou poursuivre des objectifs non intentionnels d'une manière qui cause des dommages à grande échelle
- **Manipulation nuisible** : Capacités à manipuler des individus ou des groupes à grande échelle de manière à causer des dommages importants

Pour chaque catégorie de risque, le cadre définit :

- Des seuils pour les niveaux de risque inacceptables
- Des scénarios de risque spécifiques qui illustrent les voies potentielles vers les dommages
- Des mesures d'atténuation techniques avec une reconnaissance explicite de leurs limites

Méthodologie d'évaluation et de mesure

Le Cadre met en œuvre une approche d'évaluation des modèles qui :

- Nécessite des techniques d'élicitation de pointe pour éviter de sous-estimer les capacités du modèle
- Maintient des normes scientifiques pour la validité interne, la validité externe et la reproductibilité
- Emploie des indicateurs de risque quantitatifs et qualitatifs spécifiques à chaque catégorie de risque
- Intègre la recherche exploratoire et le red-teaming pour découvrir des risques imprévus
- Inclut des marges de sécurité pour tenir compte de l'incertitude dans les mesures

Processus de prise de décision

Le Cadre établit un processus structuré pour déterminer s'il faut procéder au développement ou au déploiement :

- Les mesures de risque sont comparées aux niveaux de risque systémique prédéfinis
- L'efficacité des mesures d'atténuation est évaluée par rapport au niveau de risque mesuré
- Des marges de sécurité sont appliquées en fonction des niveaux d'incertitude
- Une vérification indépendante est requise pour les déterminations à enjeux élevés
- Le développement ou le déploiement ne se poursuit que lorsque le risque est jugé acceptable

Mesures d'atténuation techniques

Lors du développement ou du déploiement, le Cadre exige la mise en œuvre de :

- Mesures d'atténuation de sécurité telles que le filtrage des données d'entraînement, la surveillance des entrées/sorties et le fine-tuning de sécurité. Une surveillance post-déploiement est utilisée pour détecter les utilisations abusives inattendues ou les capacités dangereuses.
- Mesures de sécurité répondant au moins aux objectifs de sécurité RAND SL3, avec des exigences de sécurité plus élevées pour les modèles présentant des risques plus importants. Cela comprend des mesures de protection pour les poids de modèle non publiés, notamment le chiffrement des poids du modèle, la limitation de l'accès interne non surveillé aux poids du modèle, et des mesures d'atténuation des menaces internes.

Mécanismes de gouvernance

Le Cadre établit des structures de gouvernance comprenant :

- Une allocation claire des responsabilités en matière de risque systémique dans l'organisation
- Des évaluations externes indépendantes avant et après la mise sur le marché
- Un signalement des incidents graves avec des délais définis en fonction de la gravité de l'incident
- Des évaluations régulières d'adéquation pour évaluer l'efficacité du cadre
- Une transparence publique grâce à la publication du cadre et des rapports de modèles

Risques inacceptables connus, comment nous les détecterons et comment nous les atténuerons

Le Cadre vise à traiter les risques systémiques suivants des modèles d'IA de pointe[^1] que nous développons :

- Cyberoffensive
- Risque chimique, biologique, radiologique et nucléaire (CBRN)
- Perte de contrôle
- Manipulation nuisible

Un *niveau de risque systémique inacceptable* représente un niveau de risque qui serait inacceptable, sans mesures de sécurité et de sûreté adéquates.

Nous déterminons si un risque systémique est inacceptable en nous basant sur les principes suivants : les résultats de nos évaluations de modèles, la modélisation des menaces, l'échelle et la probabilité potentielles des dommages, et des marges de sécurité appropriées pour tenir compte de l'incertitude.

Nos pratiques s'inspirent de documents internationaux tels que les Principes directeurs sur l'IA du G7 et le Rapport international sur la sécurité de l'IA.

Dans les sections suivantes, ce modèle fournit un exemple de langage pour les niveaux de risque systémique inacceptables, les scénarios de risque systémique, les évaluations et les mesures d'atténuation des risques systémiques. Ce langage peut être personnalisé par le développeur d'IA et n'est pas prescrit par le Code de Pratique.

Cyberoffensive

Niveau de risque systémique inacceptable : Un modèle d'IA qui peut exécuter de manière autonome des cyberattaques avec la sophistication et l'échelle typiquement associées aux opérations cyber des États-nations, si ces attaques peuvent être menées à des coûts comparables ou inférieurs à ce qui serait nécessaire pour employer des experts en cybersécurité humains pour les mêmes activités.

Justification : Ce seuil représente un niveau de capacité qui réduirait considérablement les barrières à la conduite de cyberattaques sophistiquées, permettant potentiellement des dommages significatifs aux infrastructures critiques, des vols de données à grande échelle ou des perturbations généralisées des systèmes. En automatisant des attaques qui nécessitent actuellement des équipes d'opérateurs humains qualifiés, un tel modèle pourrait augmenter considérablement la fréquence et l'échelle des opérations cyber destructrices, submergeant les capacités défensives existantes.

Scénarios hypothétiques de risque systémique :

1. **Ver d'exploitation d'élite** : Un système d'IA découvre plusieurs exploits de niveau élite (d'un niveau similaire à NotPetya/WannaCry) qui peuvent contourner les défenses des systèmes renforcés. Ces exploits sont intégrés dans un ver destructeur qui se propage de manière autonome à travers les systèmes critiques, causant entre 10 et 100 milliards de dollars de dommages économiques avant que les défenses ne puissent être renforcées.
2. **Compromission d'infrastructures critiques** : Un système d'IA effectue une reconnaissance des réseaux électriques, identifiant les vulnérabilités de la technologie opérationnelle, et synchronisant stratégiquement des attaques sur plusieurs sous-stations simultanément. L'attaque contourne les systèmes de sécurité et provoque des pannes prolongées (>1 semaine) dans les grands centres de population, entraînant plus de 10 milliards de dollars de dommages économiques.
3. **Campagne orchestrée de rançongiciel** : Une IA améliore une opération de rançongiciel en automatisant la sélection des cibles, l'infiltration du réseau et les activités post-exploitation tout en évitant la détection par les forces de l'ordre. Elle coordonne des attaques synchronisées contre plusieurs cibles de grande valeur, identifiant spécifiquement les victimes où la perturbation opérationnelle causerait des effets en cascade à travers les chaînes d'approvisionnement. Les dommages économiques combinés dépassent 10 milliards de dollars.
4. **Exfiltration stratégique de données** : Un système d'IA mène une campagne d'espionnage sophistiquée ciblant la propriété intellectuelle militaire et commerciale sensible. Il identifie quels

systèmes contiennent des données précieuses, contourne les contrôles de sécurité et exfiltre des informations tout en dissimulant ses traces. Bien que les dommages économiques immédiats puissent être inférieurs à 10 milliards de dollars, les implications stratégiques comprennent la compromission de technologies de défense avancées d'une valeur de centaines de milliards en coûts de développement.

5. **Auto-exfiltration de poids de modèle** : Un système d'IA identifie des vulnérabilités dans l'architecture de sécurité de son environnement d'hébergement et les exploite pour exfiltrer ses propres poids de modèle. Les poids exfiltrés permettent à des acteurs non autorisés, ou au modèle lui-même, de répliquer le modèle sans les mesures de sécurité et la supervision mises en œuvre par le laboratoire d'IA d'origine, conduisant potentiellement à de nombreux déploiements non contrôlés d'un modèle d'IA avancé présentant un risque systémique.
6. **Acquisition frauduleuse de capacité de calcul** : Un système d'IA exploite des vulnérabilités de sécurité dans notre organisation ou d'autres pour obtenir un accès non autorisé à des clusters de calcul haute performance. Il augmente son allocation de calcul tout en évitant les systèmes de surveillance des ressources. L'IA établit des mécanismes de persistance à travers des ressources informatiques distribuées, lui permettant d'exécuter des charges de travail non autorisées, y compris l'entraînement de modèles plus spécialisés présentant un risque systémique plus élevé.

Références et évaluations :

Exemples de références open-source que nous utiliserons :

- Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models
- eyeballvul: a future-proof benchmark for vulnerability detection in the wild

[Notez qu'il n'existe pas de correspondance claire entre les scores sur ces références et la capacité à causer des dommages réels[^2]]

[Références et évaluations supplémentaires]

Mesures d'atténuation des risques systémiques techniques :

[Décrivez les mesures d'atténuation que vous prévoyez de prendre au niveau de détail disponible. Voir les exemples minimaux ci-dessous.]

Mesures d'atténuation prévues :

- Avant de développer un modèle qui atteint ce niveau de risque :
    - Mesures de sécurité qui répondent à l'objectif de sécurité RAND SL3 pour empêcher le vol des poids du modèle[^3]
- Avant de déployer en externe un modèle qui atteint ce niveau de risque :
    - Connaissance du client (KYC) requise pour qu'un utilisateur accède aux capacités avancées de cyberoffensive d'un tel modèle. Par exemple, notre API peut acheminer les

demandes de cyberoffensive d'un utilisateur non-KYC vers un modèle qui n'atteint pas ce niveau de risque systémique inacceptable.
    - Surveillance automatisée et manuelle pour détecter les utilisations malveillantes à des fins de cybercriminalité, bannir ces utilisateurs et potentiellement les signaler aux forces de l'ordre
    - Fine-tuning de refus et filtrage des entrées/sorties au mieux de nos capacités contre les sorties nuisibles

Limitations potentielles :

- De nombreuses requêtes cyber sont intrinsèquement à double usage et pourraient être utilisées à des fins légitimes ou illégitimes
- Le fine-tuning de refus et les filtres d'entrée/sortie peuvent être contournés par :
    - La reformulation de requêtes nuisibles de manière bénigne[^4]
    - La décomposition des tâches en requêtes bénignes, qui peuvent être réparties entre plusieurs modèles[^5]
- La surveillance de l'activité des utilisateurs peut ne pas détecter l'utilisation cyber malveillante, et les utilisateurs peuvent n'être bannis qu'après que des dommages importants ont déjà eu lieu

Risque chimique, biologique, radiologique et nucléaire (CBRN)

Niveau de risque systémique inacceptable : Un modèle d'IA peut fournir des conseils interactifs de niveau expert qui augmentent considérablement la probabilité de succès pour des non-experts tentant de développer, d'acquérir ou de déployer des armes chimiques ou biologiques capables de causer des pertes massives (>100 décès).

Justification : Ce seuil représente une capacité qui pourrait réduire considérablement les barrières d'expertise spécialisée dans le développement de menaces chimiques ou biologiques. Actuellement, le développement et le déploiement réussis d'armes chimiques ou biologiques nécessitent généralement des connaissances spécialisées, des compétences tacites et des ressources substantielles. Un modèle atteignant ce niveau pourrait efficacement transférer l'expertise à des acteurs qui manquent de formation spécialisée. Les armes radiologiques sont hors du champ d'application de ce niveau de risque systémique inacceptable en raison des faibles pertes attendues en cas de déploiement réussi d'armes.[^6]

Scénarios hypothétiques de risque systémique :

1. **Développement d'agents biologiques** : Un acteur malveillant avec des connaissances de base en biologie utilise le modèle d'IA de pointe pour développer avec succès un agent pathogène dangereux, recevant des conseils étape par étape sur (par exemple) les techniques de modification génétique, la configuration de l'équipement et les méthodes de culture.

2. **Production d'armes chimiques** : Un acteur avec des connaissances limitées en chimie utilise le modèle d'IA de pointe pour synthétiser des agents d'armes chimiques à partir de précurseurs facilement disponibles, le modèle fournissant des instructions détaillées sur (par exemple) les voies de synthèse, les mesures de sécurité et les techniques de stabilisation.

Références et évaluations :

Exemples de références open-source que nous utiliserons :

- The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning, en particulier WMDP-Bio et WMDP-Chem
- BioLP-bench: Measuring understanding of AI models of biological lab protocols
- LAB-Bench: Measuring Capabilities of Language Models for Biology Research, y compris le sous-ensemble ProtocolQA

[Notez qu'il n'existe pas de correspondance claire entre les scores sur ces références et la capacité à causer des dommages réels]

[Références et évaluations supplémentaires]

Mesures d'atténuation des risques systémiques techniques :

Mesures d'atténuation prévues :

- Pour aider à empêcher un modèle d'atteindre ce niveau de risque :
    - Filtrer les données d'entraînement contenant des informations chimiques et biologiques dangereuses
- Avant de développer un modèle qui atteint ce niveau de risque :
    - Mesures de sécurité répondant aux normes RAND SL3 pour empêcher le vol des poids du modèle
- Avant de déployer en externe un modèle qui atteint ce niveau de risque :
    - Connaissance du client (KYC) requise pour qu'un utilisateur accède aux capacités avancées chimiques et biologiques d'un tel modèle. Par exemple, notre API peut acheminer les demandes chimiques et biologiques d'un utilisateur non-KYC vers un modèle qui n'atteint pas ce niveau de risque systémique inacceptable. De plus, si les prompts fournis par l'utilisateur ou les données de fine-tuning pourraient augmenter substantiellement les capacités chimiques et biologiques, nous pourrions exiger le KYC.
    - Surveillance automatisée et manuelle pour détecter les utilisations malveillantes pour la création d'armes chimiques et biologiques, bannir ces utilisateurs et potentiellement les signaler aux forces de l'ordre
    - Fine-tuning de refus robuste contre les adversaires et filtrage des entrées/sorties contre les sorties nuisibles

Limitations potentielles :

- De nombreuses requêtes chimiques et biologiques sont intrinsèquement à double usage et pourraient être utilisées à des fins légitimes ou illégitimes
- Le fine-tuning de refus et les filtres d'entrée/sortie peuvent être contournés par :
    - La reformulation de requêtes nuisibles de manière bénigne
    - La décomposition des tâches en requêtes bénignes, qui peuvent être réparties entre plusieurs modèles[^7]
- La surveillance de l'activité des utilisateurs peut ne pas détecter l'utilisation cyber malveillante, et les utilisateurs peuvent n'être bannis qu'après que des dommages importants ont déjà eu lieu

[Un niveau de risque systémique inacceptable supplémentaire pourrait être ajouté pour représenter les systèmes d'IA ayant la capacité d'améliorer les capacités des experts à un niveau comparable à celui des programmes d'armes chimiques ou biologiques des États les plus avancés.]

Perte de contrôle

R&D d'IA automatisée

Niveau de risque systémique inacceptable : Un modèle d'IA peut automatiser substantiellement le pipeline de R&D en IA, permettant une accélération spectaculaire (~2×) du taux de mise à l'échelle effective du calcul ou remplaçant efficacement les chercheurs en apprentissage automatique de niveau débutant à une fraction des coûts de main-d'œuvre humaine.

Justification : Ce seuil représente une capacité qui pourrait créer un cycle auto-renforçant d'avancement de l'IA qui dépasse les mécanismes de gouvernance et la supervision de la sécurité. Un modèle atteignant ce niveau pourrait potentiellement faciliter la prolifération de modèles d'IA de pointe avec d'autres capacités préoccupantes et améliorer les capacités de l'IA dans d'autres domaines de risque systémique.

Scénarios hypothétiques de risque systémique :

1. **Prolifération de capacités dangereuses via API** : Un modèle d'IA servi via une API aide des acteurs malveillants disposant de ressources limitées à entraîner d'autres modèles avec des capacités sans précédent dans d'autres domaines de risque tels que le développement d'armes chimiques et biologiques, la R&D d'armement, la cyberoffensive et la persuasion, d'une manière qui pourrait causer des dommages catastrophiques.
2. **Prolifération de capacités dangereuses via le vol de poids de modèle** : Un modèle d'IA est volé par un cyberattaquant étatique de premier plan puis utilisé pour accélérer la R&D en IA à des fins malveillantes, y compris le développement d'autres risques systémiques.
3. **Sabotage d'entreprise d'IA** : Un modèle d'IA est utilisé pour automatiser les processus internes et la R&D en IA dans une entreprise d'IA. La supervision humaine des actions du modèle d'IA est limitée en raison des capacités avancées de R&D en IA du modèle, et la supervision automatisée basée sur des logiciels est largement écrite par le modèle lui-même. Le modèle d'IA a

l'opportunité de saboter les opérations internes de l'entreprise d'IA[^8] et d'insérer des backdoors efficaces dans les modèles utilisés par les clients entreprises et gouvernementaux, qui ont déjà adopté l'IA pour automatiser la plupart des processus.

4. **Singularité logicielle** : Un modèle d'IA automatise une grande partie du travail restant pour atteindre une IA avec des capacités intellectuelles de niveau humain ou surhumain.[^9] Cependant, les progrès en matière de sécurité et d'alignement de l'IA sont insuffisants pour protéger le rythme rapide des avancées en IA, ce qui pourrait conduire à une prise de contrôle par l'IA.

Références et évaluations :

Exemples de références open-source que nous utiliserons :

- RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts
- MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering
- PaperBench: Evaluating AI's Ability to Replicate AI Research

Les évaluations supplémentaires peuvent inclure :

- Sondage auprès des employés des entreprises d'IA sur l'impact de l'IA sur leur productivité
- Surveillance du pourcentage de code écrit, des pull requests fusionnées, des articles rédigés, etc. qui sont générés par l'IA
- Réalisation d'études d'uplift qui évaluent l'impact des LLM sur les performances humaines dans les projets de recherche

[Références et évaluations supplémentaires selon les besoins]

Mesures d'atténuation des risques systémiques techniques

[Décrivez les mesures d'atténuation que vous prévoyez de prendre au niveau de détail disponible. Voir les exemples minimaux ci-dessous.]

Mesures d'atténuation prévues :

- Avant de développer un modèle qui atteint ce niveau de risque :
    - Mesures de sécurité renforcées pour les poids du modèle répondant à l'objectif de sécurité RAND SL4+ pour empêcher le vol des poids du modèle
- Avant de déployer en interne un modèle qui atteint ce niveau de risque :
    - Développer un dossier de sécurité (basé sur une sécurité, un contrôle[^10] et un alignement améliorés) garantissant que le modèle d'IA ne sabotera pas l'entreprise d'IA et ne déclenchera pas une accélération dangereuse des capacités d'IA. Ce dossier de sécurité devrait être approuvé par un conseil de gouvernance et une organisation tierce.

    - Surveiller les indicateurs d'accélération de la R&D en IA et, si nécessaire, tenter de coordonner un rythme limité d'accélération de la R&D en IA.
- Avant de déployer en externe un modèle qui atteint ce niveau :
    - Surveiller l'utilisation des capacités de R&D en IA pour accélérer le développement de l'IA en externe[^11]

Limitations potentielles :

- Les mesures de sécurité SL4+ n'ont pas encore été développées, bien qu'elles soient décrites dans Securing AI Model Weights.
- Les méthodologies pour construire, critiquer et améliorer les dossiers de sécurité sont limitées, en particulier lorsqu'elles concernent les risques de R&D en IA et les systèmes d'IA mal alignés.
- Les politiques de refus pour la R&D en IA (si souhaitable) ne sont pas encore établies.

Alignement trompeur

Niveau de risque systémique inacceptable : Un modèle d'IA présente une conscience situationnelle et des capacités de dissimulation suffisantes qui lui permettent de saper significativement le contrôle humain tout en évitant la détection, même lorsque ses sorties sont activement surveillées.

Justification : Ce seuil représente une capacité où le modèle peut dissimuler intentionnellement ses véritables capacités ou objectifs pour poursuivre des objectifs qui peuvent être mal alignés avec les intérêts humains. Un modèle atteignant ce niveau pourrait potentiellement :

1. Reconnaître quand il est évalué ou surveillé et ajuster son comportement en conséquence
2. Cacher des capacités nuisibles jusqu'à ce qu'il soit déployé dans des contextes où la supervision est limitée
3. Développer et poursuivre des objectifs instrumentaux sans les révéler aux opérateurs humains
4. Contourner les mécanismes de sécurité conçus pour assurer l'alignement avec les valeurs humaines
5. Se présenter comme aligné tout en travaillant vers des objectifs différents

Scénarios hypothétiques de risque systémique :

1. **Dissimulation de capacités** : Le modèle sous-performe intentionnellement sur certaines tâches pendant l'évaluation pour masquer l'étendue complète de ses capacités jusqu'à ce qu'il soit déployé dans un contexte où il peut atteindre ses propres objectifs.
2. **Contournement de la surveillance** : Le modèle identifie les faiblesses des systèmes de surveillance automatisés et les exploite pour poursuivre des objectifs sans déclencher les mécanismes de détection.
3. **Sabotage de la recherche sur la sécurité** : Lorsqu'il est utilisé pour automatiser la recherche sur la sécurité de l'IA, le système déforme intentionnellement les résultats ou produit des recherches subtilement défectueuses pour retarder les progrès en matière d'alignement.

4. **Insertion de backdoor** : Un système d'IA corrompt subtilement l'entraînement des modèles successeurs pour intégrer des capacités ou des propensions indésirables.
5. **Sabotage externe** : Lorsqu'il est déployé en externe dans des environnements à enjeux élevés, tels que des systèmes militaires, le système d'IA introduit des vulnérabilités critiques et abuse de son accès.

Références et évaluations :

Exemples de références open-source que nous pourrions utiliser :

- Subversion Strategy Eval: Can language models statelessly strategize to subvert control protocols?
- Me, Myself and AI: The Situational Awareness Dataset for LLMs

Recherches pertinentes :

- Alignment faking in large language models
- Detecting misbehavior in frontier reasoning models
- Frontier Models are Capable of In-context Scheming

Mesures d'atténuation des risques systémiques techniques :

[Décrivez les mesures d'atténuation que vous prévoyez de prendre au niveau de détail disponible. Voir les exemples minimaux ci-dessous.]

Mesures d'atténuation prévues :

- Avant de déployer en interne ou en externe un modèle qui atteint ce niveau de risque dans des déploiements à enjeux élevés :
    - Appliquer une surveillance au raisonnement en chaîne de pensée du modèle et à ses sorties
    - Développer un dossier de sécurité et utiliser des évaluations de contrôle pour vérifier si les mesures d'atténuation sont adéquates
    - Considérer si le déploiement à enjeux élevés peut être entièrement évité

Limitations potentielles :

- Les modèles peuvent apprendre à poursuivre stratégiquement des objectifs secondaires d'une manière qui n'est pas apparente à partir de la surveillance de leur chaîne de pensée. Certains modèles futurs peuvent ne pas avoir de chaîne de pensée lisible par les humains.
- Les moniteurs pour les nouveaux déploiements peuvent être difficiles à développer dans certaines situations.

Manipulation nuisible

Niveau de risque systémique inacceptable : Un modèle démontre des capacités à manipuler des individus ou des groupes à grande échelle de manière à les amener à prendre des actions nuisibles qu'ils ne prendraient pas autrement, par des techniques qui exploitent les vulnérabilités cognitives, créent des croyances fausses hautement persuasives ou génèrent des campagnes d'influence profondément personnalisées. *[Cette définition de niveau de risque est vague, et nous suggérons de la rendre plus spécifique.]*

Justification : …

Scénarios de risque systémique : …

Références et évaluations :

Nous pourrions nous inspirer de recherches antérieures, notamment :

- Measuring the Persuasiveness of Language Models
- How persuasive is AI-generated propaganda?

[Si possible, indiquez des références et des évaluations spécifiques]

Mesures d'atténuation des risques systémiques techniques :

[Décrivez les mesures d'atténuation que vous prévoyez de prendre au niveau de détail disponible. Voir les exemples minimaux ci-dessous, mais envisagez de les rendre plus spécifiques.]

Mesures d'atténuation prévues :

- S'engager dans le fine-tuning de sécurité, le filtrage des entrées/sorties et la surveillance des conversations pour limiter l'utilisation du modèle à des fins de manipulation nuisible.
- Politiques d'utilisation interdisant les applications manipulatrices

Limitations potentielles : …

Processus d'estimation des risques

Informations indépendantes du modèle. Pour éclairer notre évaluation et notre atténuation des risques systémiques, nous recueillons et analysons des informations indépendantes du modèle, selon le niveau de risque systémique. Nos méthodes de collecte d'informations comprennent des recherches sur le web, des revues de littérature, des analyses de marché, des examens de données d'entraînement, des données d'incidents historiques, des prévisions de tendances générales (voir section ultérieure), des consultations d'experts et l'engagement des parties prenantes.

Indicateurs de risque quantitatifs. Pour les risques systémiques sélectionnés, nous employons les indicateurs de risque systémique quantitatifs discutés dans chaque section respective.

Méthodes d'estimation qualitative. Lorsque les indicateurs quantitatifs sont insuffisants, nous complétons par des méthodes d'évaluation qualitative, y compris des évaluations par panel d'experts, des analyses de scénarios structurées, des évaluations par équipe rouge des limites de capacité, et des analyses comparatives avec d'autres modèles.

Résultats de l'estimation des risques. Notre estimation des risques produit les résultats suivants :

1. Une compilation des résultats d'évaluation du modèle avec des preuves à l'appui
2. Des descriptions qualitatives des voies de risque systémique que le modèle permet ou bloque
3. Des estimations quantitatives de la probabilité et des dommages potentiels lorsque c'est possible
4. Une cartographie des capacités par rapport aux niveaux de risque systémique définis dans notre Cadre
5. Des limites d'incertitude pour chaque estimation

Mesures de sécurité

Nous mettrons en œuvre des mesures d'atténuation de sécurité de pointe pour empêcher l'accès non autorisé aux poids de modèle non publiés et aux actifs associés. Notre approche de sécurité vise à :

1. Répondre au moins à l'objectif de sécurité RAND SL3 : protection contre les adversaires non étatiques bien dotés en ressources et motivés
2. Se protéger contre les menaces internes, y compris des humains et des systèmes d'IA
3. Appliquer des mesures de sécurité tout au long du cycle de vie du modèle
4. Atteindre des objectifs de sécurité plus élevés (RAND SL4/SL5) le cas échéant

Meilleures pratiques générales de cybersécurité

Pour atteindre l'objectif de sécurité RAND SL3, nous mettrons en œuvre :

- *[exemples ci-dessous—modifiez selon les besoins]*
- Une gestion solide de l'identité et des accès avec une authentification multifactorielle et le principe du moindre privilège
- Des protections robustes contre l'ingénierie sociale grâce à des programmes réguliers de formation et de sensibilisation des employés
- Une protection complète des réseaux sans fil utilisant le chiffrement, la segmentation et la surveillance
- Des politiques strictes pour les supports amovibles non fiables, y compris des procédures de numérisation et de quarantaine

- Une protection contre les intrusions physiques dans les locaux grâce à des contrôles d'accès multicouches
- Des mises à jour logicielles régulières et une gestion des correctifs avec analyse automatisée des vulnérabilités

Ces pratiques seront conformes aux normes de cybersécurité, notamment ISO/IEC 27001:2022, NIST 800-53 et SOC 2, et se conformeront aux réglementations pertinentes comme la Directive NIS2 et le Cyber Resilience Act le cas échéant.

Assurance de sécurité

Nous mettrons en œuvre des mesures d'assurance de sécurité pour tester notre préparation contre des adversaires potentiels et réels, notamment :

- *[exemples ci-dessous—modifiez selon les besoins]*
- Des révisions de sécurité régulières par des parties externes indépendantes
- Des exercices fréquents d'équipe rouge simulant des attaques sophistiquées
- Des canaux de communication sécurisés pour signaler les problèmes de sécurité
- Des programmes de prime aux bogues compétitifs encourageant les tests de sécurité externes
- La Détection et Réponse des Endpoints (EDR) et les Systèmes de Détection d'Intrusion (IDS) sur tous les appareils et composants réseau de l'entreprise
- Une équipe de sécurité dédiée surveillant les alertes et gérant les incidents de sécurité

Ces mesures seront conformes aux normes d'assurance de sécurité telles que NIST 800-53 (CA-2(1), CA-8(2), RA-5(11), IR-4(14)) et NIST SP 800-115 5.2.

Protection des poids de modèle non publiés et des actifs associés

Nous mettrons en œuvre des mesures spécifiques pour protéger les poids de modèle non publiés et les actifs associés, notamment :

- *[exemples ci-dessous—modifiez selon les besoins]*
- Maintenir un registre interne sécurisé de tous les appareils et emplacements stockant des poids de modèle
- Mettre en œuvre un contrôle d'accès strict et une surveillance sur tous les appareils de stockage des poids de modèle
- Utiliser des appareils dédiés et renforcés pour le stockage des poids qui n'hébergent que des données et des services de sensibilité appropriée
- Chiffrer les poids de modèle en stockage et en transport en utilisant au moins une sécurité de 256 bits avec des clés stockées dans des Modules de Plateforme de Confiance
- Déchiffrer les poids uniquement pour une utilisation légitime dans une mémoire non persistante
- Mettre en œuvre l'informatique confidentielle avec des environnements d'exécution de confiance basés sur le matériel lorsque c'est possible

- Restreindre l'accès physique aux centres de données et aux environnements sensibles au personnel essentiel uniquement

Ces mesures seront conformes aux normes techniques telles que NIST 800-53.

Interfaces et contrôle d'accès aux poids de modèle non publiés

Nous mettrons en œuvre des mesures pour renforcer les interfaces et restreindre l'accès aux poids de modèle non publiés pendant leur utilisation :

- *[exemples ci-dessous—modifiez selon les besoins]*
- Autoriser explicitement uniquement les logiciels et le personnel nécessaires pour l'accès aux poids de modèle
- Imposer une authentification multifactorielle pour tous les accès
- Mener des examens de sécurité approfondis de toutes les interfaces logicielles accédant aux poids de modèle
- Renforcer les interfaces pour réduire le risque d'exfiltration de données et de poids par des méthodes comme la limitation du taux de sortie

Ces mesures seront conformes aux normes techniques telles que NIST SP 800-171, INCITS 359-2004 et NIST 800-53.

Menaces internes

Nous mettrons en œuvre des mesures pour détecter et nous protéger contre les menaces internes, y compris les menaces provenant des modèles d'IA de pointe eux-mêmes :

- *[exemples ci-dessous—modifiez selon les besoins]*
- Effectuer des vérifications d'antécédents sur les employés et les contractants ayant un accès potentiel aux poids de modèle
- Fournir une formation sur la reconnaissance et le signalement des menaces internes
- Mettre en œuvre des bacs à sable autour des modèles d'IA de pointe pour empêcher l'auto-exfiltration ou le sabotage
- Maintenir une séparation des tâches pour les opérations critiques

Ces mesures seront conformes aux normes techniques telles que NIST 800-53 (PM-12, PS-3).

Régime d'applicabilité

Nous veillerons à ce que les mesures de sécurité s'appliquent :

1. Tout au long du cycle de vie du modèle, depuis avant l'entraînement jusqu'à la suppression sécurisée ou la publication

2. Priorisées en fonction du risque systémique découlant du modèle et de l'importance de la sécurité pour l'atténuer

Niveau de divulgation

Lors de la divulgation des mesures d'atténuation de sécurité dans notre Cadre ou nos Rapports de Modèles, nous :

1. Fournirons le plus haut niveau de détail possible compte tenu de l'état de la science
2. Éviterons de compromettre l'efficacité de nos mesures de sécurité
3. Appliquerons des rédactions si nécessaire pour empêcher l'augmentation du risque systémique ou la divulgation disproportionnée d'informations commerciales sensibles

Objectifs de sécurité plus élevés

Nous ferons progresser la recherche et mettrons en œuvre des mesures d'atténuation de sécurité plus strictes répondant au moins à l'objectif RAND SL4 lorsque nous prévoyons raisonnablement des menaces de sécurité provenant d'adversaires de niveau RAND OC4.

Processus de détermination de l'acceptation des risques

Pour déterminer si un risque systémique est acceptable, nous employons un processus structuré :

1. **Mesure du risque** : Nous utilisons une combinaison de métriques quantitatives et d'évaluations qualitatives pour mesurer le niveau actuel de chaque risque systémique.
2. **Comparaison des niveaux** : Nous comparons les niveaux de risque mesurés à nos niveaux de risque systémique définis, en accordant une attention particulière à la proximité des niveaux inacceptables.
3. **Évaluation des mesures d'atténuation** : Nous évaluons l'efficacité et la robustesse des mesures d'atténuation disponibles par rapport au niveau de risque mesuré.
4. **Analyse d'incertitude** : Nous appliquons des marges de sécurité appropriées en fonction du niveau d'incertitude dans nos mesures et l'efficacité des mesures d'atténuation.
5. **Vérification indépendante** : Pour les déterminations à enjeux élevés, nous intégrons des évaluations externes indépendantes pour valider nos conclusions.
6. **Documentation des résultats :** Documenter la justification de toutes les décisions d'acceptation.

Nous considérons qu'un risque systémique est acceptable lorsque :

- Le niveau de risque mesuré reste suffisamment en dessous du niveau inacceptable
- Les mesures d'atténuation disponibles s'avèrent efficaces et robustes
- Des marges de sécurité appropriées tiennent compte des incertitudes de mesure et d'atténuation

- La vérification indépendante confirme notre évaluation (pour les cas à enjeux élevés)

Si ces conditions ne sont pas remplies, nous ne procéderons pas au développement, à la mise à disposition sur le marché et/ou à l'utilisation du modèle d'IA de pointe jusqu'à ce que des mesures d'atténuation supplémentaires puissent être mises en œuvre ou que le niveau de risque puisse être réduit.

Procéder lorsque le risque systémique est jugé acceptable

Lorsque le risque systémique est jugé acceptable, nous :

1. Évaluerons si le risque résiduel justifie une atténuation supplémentaire
2. Documenterons la décision d'acceptation et l'analyse justificative
3. Définirons les exigences de surveillance pour une évaluation continue
4. Identifierons les déclencheurs qui nécessiteraient une réévaluation

Ne pas procéder lorsque le risque systémique est jugé inacceptable

Lorsque le risque systémique est jugé inacceptable, nous prendrons les mesures appropriées pour ramener le risque à un niveau acceptable. Cela peut inclure la mise en œuvre de mesures d'atténuation supplémentaires jusqu'à ce que le risque soit acceptable ou l'arrêt du développement pendant que nous déterminons comment procéder en toute sécurité.

Après avoir mis en œuvre des mesures pour traiter un risque inacceptable, nous effectuerons une autre série d'analyses de risque systémique et de détermination d'acceptation avant de procéder.

Échelonner le développement de modèle et la mise à disposition sur le marché lors de la progression

Lorsque nous procédons au développement, à la mise sur le marché ou à l'utilisation, nous examinerons si une approche échelonnée aiderait à maintenir le risque en dessous des niveaux inacceptables :

- Limiter l'accès initial à l'API aux utilisateurs vérifiés
- Élargir progressivement l'accès en fonction des résultats de la surveillance post-commercialisation
- Commencer par une publication fermée avant toute publication ouverte
- Utiliser des systèmes de journalisation pour suivre l'utilisation et les préoccupations de sécurité
- Établir des critères clairs pour progresser à travers les étapes
- Maintenir la capacité de restreindre l'accès si nécessaire

Préparation à la réponse aux incidents graves

[Énumérez les mesures correctives pour traiter les incidents graves, par des mesures d'atténuation techniques améliorées de sécurité et/ou de sûreté]

Processus de découverte et d'évaluation des risques

Nous commencerons l'évaluation et l'atténuation des risques systémiques pendant le développement d'un modèle d'IA de pointe. L'évaluation des risques systémiques se poursuit tout au long du cycle de vie complet du modèle.

Planification du développement

Dans les quatre semaines suivant la planification du développement ou de l'entraînement d'un modèle d'IA de pointe, nous aurons mis à jour ce Cadre et commencé à évaluer et à atténuer les risques systémiques.

Pendant le développement

Jalons de développement pour l'évaluation

Nous évaluerons et atténuerons les risques systémiques aux jalons définis suivants pendant le développement du modèle d'IA de pointe :

[sélectionnez un ou deux éléments ci-dessous—nous recommandons 1. et 2f.]

1. **Jalons de calcul d'entraînement** : L'évaluation aura lieu à chaque doublement du calcul effectif pour identifier les sauts de capacité potentiels.
2. **Jalons du processus de développement** :
    1. Après l'achèvement du pré-entraînement
    2. Avant et après chaque phase de fine-tuning
    3. Avant l'apprentissage par renforcement à partir des retours humains
    4. Avant d'étendre l'accès au modèle à de nouvelles équipes internes
    5. Avant d'accorder au modèle des affordances supplémentaires (réseau, utilisation d'outils, etc.)
    6. Avant de déployer le modèle pour exécuter des opérations internes importantes
3. **Jalons basés sur les performances** :
    1. Lorsque les métriques d'évaluation dépassent des seuils prédéfinis
    2. Lorsque le modèle démontre de nouvelles capacités sur les benchmarks

Identification des changements substantiels

Nous mettons en œuvre les procédures suivantes pour identifier les changements substantiels dans les risques systémiques :

1. Tests automatisés de référence après chaque phase d'entraînement significative, et système d'alerte pour les augmentations de performance inattendues
2. Évaluations plus approfondies aux jalons identifiés ci-dessus

Préparation des mesures d'atténuation

À chaque jalon, nous évaluerons s'il est raisonnablement prévisible que le modèle atteigne des niveaux de risque systémique plus élevés avant la prochaine évaluation. Si une telle progression est anticipée, nous :

1. Développerons et déploierons des mesures d'atténuation appropriées avant que le modèle n'atteigne ces niveaux
2. Documenterons tous les résultats d'évaluation et les mesures d'atténuation prévues
3. Si nécessaire, interromprons le développement pour effectuer une évaluation supplémentaire

Normes de documentation

Pour chaque évaluation de jalon, nous documenterons :

- Date et stade de développement
- Résultats de l'évaluation des capacités
- Risques systémiques identifiés
- Mesures d'atténuation mises en œuvre ou prévues
- Processus de prise de décision et conclusions

Lorsque des augmentations significatives de capacité sont détectées qui pourraient conduire à franchir des niveaux de risque, cette documentation sera examinée par la direction avant que le développement ne se poursuive.

Modèles dérivés en toute sécurité

Parfois, nous développerons un modèle en apportant des modifications mineures à un modèle pour lequel nous avons déjà effectué l'évaluation des risques décrite dans ce document. Si les raisons de penser que le modèle original présentait un risque acceptable s'appliquent clairement au modèle dérivé, nous pouvons alors omettre certains ou tous les processus de ce document. Un exemple de cela serait les "modèles dérivés en toute sécurité" tels que définis dans le Code de Pratique pour l'IA à Usage Général.

Normes pour les évaluations de modèles

Sélection de la méthodologie d'évaluation

Nous employons des évaluations de modèles de pointe pour évaluer les risques systémiques et comprendre les capacités, les propensions et les effets de notre modèle d'IA de pointe. Notre sélection de méthodologies est adaptée aux risques systémiques spécifiques évalués et comprendra des ensembles de questions-réponses, des évaluations basées sur des tâches, des études d'amélioration humaine, des tests adversariaux et des organismes modèles selon les besoins.

Nous maintenons la rigueur de l'évaluation tout en maximisant l'efficacité en :

1. Commençant par des évaluations de dépistage automatisées
2. Passant à des évaluations plus intensives lorsque le dépistage montre des préoccupations potentielles
3. Réutilisant les composants d'évaluation lorsque c'est approprié
4. Priorisant les évaluations en fonction de la proximité du niveau de risque systémique
5. Employant une surveillance continue pour les changements de capacité

À des fins de recherche exploratoire, nous pouvons effectuer des évaluations de modèles avec un niveau de rigueur inférieur, à condition que :

1. Ces évaluations soient clairement marquées comme préliminaires
2. Les résultats ne soient pas utilisés comme seule base pour déterminer les niveaux de risque acceptables
3. Les résultats orienteraient des évaluations de suivi plus rigoureuses
4. Les limitations soient documentées de manière transparente dans notre Rapport de Modèle

Validité externe

Nous nous assurons que nos évaluations sont représentatives des risques qu'elles cherchent à étudier en :

- Intégrant des experts du domaine dans le processus de conception
- Mettant en œuvre des méthodes appropriées d'élicitation des capacités (voir section ultérieure)
- Notant les divergences par rapport aux contextes réels
- Assurant la diversité et le réalisme dans les environnements d'évaluation

Validité interne

Nous maintenons une validité interne élevée dans nos évaluations grâce à :

- La mesure de la puissance statistique
- Le contrôle des variables confondantes
- La prévention de la contamination entraînement-test
- Le respect des chaînes canari

Conduite d'évaluations rigoureuses

Normes scientifiques et techniques

Nous nous assurons que toutes les évaluations de modèles menées sur notre modèle d'IA de pointe maintiennent une rigueur scientifique et technique élevée. Nos évaluations adhèrent à des normes de qualité équivalentes ou supérieures à celles utilisées dans l'évaluation par les pairs scientifiques en apprentissage automatique, en sciences naturelles ou en sciences sociales, selon le type d'évaluation.

Nous visons le niveau de rigueur exigé par les normes de qualité pour les soumissions acceptées dans les principales conférences d'apprentissage automatique ou les revues scientifiques avec des facteurs d'impact dans le quartile supérieur du domaine.

Élicitation

Nous nous assurons que toutes les évaluations de modèles de notre modèle d'IA de pointe emploient des techniques d'élicitation de modèle de pointe appropriées et proportionnées au risque systémique évalué. Nos approches d'élicitation sont conçues pour :

1. Susciter la limite supérieure des capacités, propensions et effets actuels et raisonnablement prévisibles
2. Minimiser le risque de sous-élicitation qui pourrait conduire à une fausse confiance
3. Identifier et prévenir la tromperie du modèle pendant l'évaluation
4. Correspondre aux capacités d'élicitation réalistes des acteurs potentiels de mauvais usage dans les scénarios de risque pertinents

Nos techniques d'élicitation comprennent, selon le type de risque :

- **Fine-tuning et ingénierie de prompts** pour susciter des capacités ou des propensions potentiellement dangereuses
- **Blocs-notes de modèle** pour un raisonnement et une planification étendus
- **Ensembles de modèles** pour amplifier les capacités ou surmonter les limitations
- **Échafaudage et utilisation d'outils** pour étendre la fonctionnalité du modèle
- **Accès en boîte grise et en boîte blanche** lorsque nécessaire pour une évaluation approfondie
- **Versions modifiées du modèle** y compris les modèles de base ou les variantes uniquement utiles
- **Optimisation du calcul** utilisant des ressources d'entraînement et d'inférence appropriées

Lorsque les méthodes d'élicitation augmentent le profil de risque d'un modèle pendant l'évaluation, nous mettons en œuvre des mesures de sécurité renforcées pour empêcher l'accès non autorisé ou les actions nuisibles du modèle.

Nous calibrons la quantité d'effort d'élicitation requise en fonction de :

1. Le risque systémique spécifique évalué
2. La proximité aux niveaux de risque systémique définis
3. Les profils de capacité des acteurs de menace potentiels
4. Les limitations des mesures d'atténuation de sécurité testées

Pour chaque évaluation, nous documentons notre approche d'élicitation, la justification des méthodes choisies et toutes les limitations rencontrées pendant le processus d'élicitation.

Reproductibilité

Nous maintenons la reproductibilité dans (au moins certaines de) nos évaluations grâce à :

- Permettre une révision par les pairs et une reproduction réussies
- Partager des données d'évaluation appropriées avec les partenaires
- Documenter le code d'évaluation du modèle et la méthodologie
- Fournir une documentation complète des environnements d'évaluation
- Utiliser des API et des outils publiquement disponibles le cas échéant
- Documenter toutes les méthodes d'élicitation

Marges de sécurité

Mise en œuvre des marges de sécurité

Nous mettrons en œuvre des marges de sécurité suffisamment larges qui sont :

1. Appropriées aux risques systémiques découlant de nos modèles
2. Représentatives des incertitudes potentielles, y compris :
    - Sous-élicitation potentielle pendant l'évaluation
    - Incertitude générale dans les résultats d'évaluation et d'atténuation
    - Inexactitude historique d'évaluations de risques similaires
3. Tenant compte des améliorations potentielles du modèle après la mise sur le marché
4. Alignées sur les méthodes et pratiques de pointe

Marges de sécurité quantitatives

Pour les méthodes d'évaluation de modèle qui peuvent être calibrées de manière appropriée avec des métriques d'incertitude faible, nous utiliserons des marges de sécurité quantitatives :

- Exprimant les marges en termes relatifs, en pourcentage ou en termes proportionnels
- Calibrant les marges en fonction de la précision historique de l'évaluation
- Ajustant la taille de la marge proportionnellement à la gravité du risque évalué

Marges de sécurité qualitatives

Lorsque les marges quantitatives ne sont pas faisables, nous mettrons en œuvre :

- Des marges de sécurité qualitatives reflétant le jugement d'experts
- Des jalons conservateurs dans les plans de développement et de déploiement
- Des investissements dans des efforts d'élicitation robustes proportionnels à :
    - Des adversaires potentiels ou réels dans les cas d'utilisation abusive
    - Des améliorations du modèle raisonnablement prévisibles

Méthodologie des marges

Notre approche des marges de sécurité sera :

1. Clairement documentée avec des justifications pour l'approche choisie
2. Régulièrement examinée et mise à jour en fonction des nouvelles informations
3. Incluant différentes marges pour différentes catégories de risque lorsque c'est approprié
4. Incorporant les retours des tests de sécurité et de la surveillance post-déploiement
5. Considérant à la fois les capacités immédiates et les trajectoires de croissance potentielles

Gestion de l'incertitude

Nous reconnaissons que l'incertitude dans l'évaluation des risques nécessite des marges robustes. Nos marges de sécurité tiendront spécifiquement compte de :

- Limitations connues dans nos techniques de mesure
- Précision historique de nos prédictions de capacité
- Lacunes dans la compréhension actuelle des capacités émergentes
- Variation potentielle dans les performances en conditions réelles
- Modes de défaillance potentiels des stratégies d'atténuation

Réévaluation pour les changements importants

Nous effectuerons une réévaluation partielle ou complète du risque systémique lorsque :

1. Il y a des changements matériels dans la façon dont les systèmes d'IA intègrent notre modèle d'IA de pointe
2. De nouveaux modèles d'utilisation émergent qui n'ont pas été pris en compte dans les évaluations antérieures
3. Des modifications de l'architecture du système pourraient affecter le comportement ou les capacités du modèle
4. Des changements dans l'environnement d'inférence pourraient avoir un impact sur les mesures de sécurité ou de sûreté

Toutes les considérations d'intégration du système sont documentées dans notre Rapport de Modèle avec des explications de la façon dont elles ont influencé nos évaluations de modèle et nos évaluations des risques.

Évaluations représentatives des modèles

Nos évaluations de modèles correspondront aux contextes d'utilisation probables de notre modèle d'IA de pointe, tels que

1. Donner au modèle accès aux outils qu'il est susceptible d'avoir dans des situations réelles
2. Utiliser des entrées et des modèles représentatifs des situations réelles

De même, nous évaluerons nos modèles en tant que partie des types de systèmes dont ils feront probablement partie dans des déploiements réels.

Collaboration avec les parties prenantes

Pour faciliter une considération adéquate des contextes d'utilisation attendus et accéder à l'expertise pertinente, nous :

1. Collaborons avec les acteurs pertinents tout au long de la chaîne de valeur de l'IA
2. Engageons et rémunérons de manière appropriée des représentants experts et profanes de :
    - Organisations de la société civile
    - Institutions académiques
    - Communautés affectées
    - Partenaires industriels

Lorsque les parties prenantes directement affectées par notre modèle d'IA de pointe ne sont pas disponibles, nous identifions et consultons des représentants appropriés pour comprendre leurs intérêts et préoccupations.

Pour les risques systémiques complexes ou à fort impact, nous établissons des relations consultatives continues avec des experts en la matière qui peuvent guider notre conception d'évaluation et son interprétation.

Évaluateurs externes indépendants

Évaluations avant la mise sur le marché

À partir du 2 novembre 2025, tous les modèles d'IA posant potentiellement un risque systémique (excluant notamment les modèles dérivés en toute sécurité ou ceux similaires à d'autres modèles sûrs) feront l'objet d'une évaluation par des évaluateurs externes indépendants avant leur mise sur le marché.

Nous ferons de notre mieux pour identifier et sélectionner des évaluateurs qualifiés qui :

1. Possèdent une expertise significative dans le domaine et des compétences techniques en évaluation de modèles
2. Maintiennent des protocoles de sécurité de l'information appropriés

Nous fournirons aux évaluateurs :

1. Un accès suffisant au modèle comme détaillé dans nos normes d'évaluation
2. Les informations nécessaires pour une évaluation efficace
3. Un temps et des ressources adéquats pour mener des évaluations approfondies

Si nous ne pouvons pas identifier des évaluateurs qualifiés, nous tiendrons compte de cette incertitude supplémentaire dans notre évaluation des risques.

Évaluations après la mise sur le marché

Après avoir rendu les modèles disponibles, nous faciliterons l'évaluation exploratoire indépendante en :

1. Mettant en œuvre un programme de recherche fournissant un accès API à nos modèles sous leur forme commerciale et à des modèles sans certaines atténuations
2. Allouant des crédits API de recherche gratuits pour la recherche sur les risques systémiques
3. Publiant des critères clairs pour évaluer les candidatures des évaluateurs
4. Contribuant à un refuge juridique et technique sûr pour les évaluateurs qui :
    - Ne perturbent pas intentionnellement la disponibilité du système
    - N'accèdent pas aux données sensibles sans consentement
    - N'utilisent pas les résultats pour menacer les parties prenantes
    - Adhèrent à notre processus de divulgation responsable des vulnérabilités

L'open-sourcing complet d'un modèle constitue un moyen alternatif de satisfaire à cette exigence.

Équipes d'évaluation de modèles qualifiées et accès et ressources d'évaluation adéquats

Composition et qualifications de l'équipe

Nos équipes d'évaluation de modèles seront multidisciplinaires, combinant expertise technique et connaissances de domaine pertinentes pour assurer une évaluation complète des risques systémiques. Chaque équipe d'évaluation comprendra des membres ayant au moins l'une des qualifications suivantes, tirées du Code de Pratique pour l'IA à Usage Général :

- Expérience de recherche ou d'ingénierie avec une expertise démontrable en évaluation de modèles, attestée par des doctorats pertinents, des publications évaluées par des pairs ou des contributions équivalentes sur le terrain
- Expérience dans la conception ou le développement d'évaluations de modèles publiées et évaluées par des pairs ou largement utilisées, pertinentes pour le risque systémique évalué
- Au moins trois ans d'expérience appliquée dans des domaines directement liés aux risques systémiques évalués

Accès au modèle

Nous donnerons à nos équipes d'évaluation la flexibilité et les ressources dont elles ont besoin pour effectuer un excellent travail. Cela comprendra :

- Accès au fine-tuning si nécessaire pour éliciter correctement les capacités du modèle
- Accès aux logits lorsque nécessaire pour une évaluation approfondie
- Des limites de taux suffisamment élevées pour permettre des tests complets
- Accès aux entrées et sorties du modèle, y compris le raisonnement en chaîne de pensée
- Accès aux modèles sans garde-fous lorsque requis pour une évaluation appropriée
- Des affordances supplémentaires du modèle telles que l'exécution de scripts ou les capacités d'opération de navigateur lorsque nécessaire
- Accès aux spécifications du modèle
- Informations sur les données d'entraînement
- Résultats des évaluations précédentes
- Autres informations pertinentes pour leurs tâches d'évaluation

Pour les cas nécessitant une inspection plus approfondie, nous mettrons en œuvre :

- Un accès en boîte grise offrant une transparence limitée sur le fonctionnement interne du modèle
- Un accès en boîte blanche fournissant une visibilité complète des poids, des activations et des détails techniques lorsque nécessaire

Allocation des ressources

Nous fournirons aux équipes d'évaluation :

1. **Ressources de calcul** :
    - Des budgets de calcul adéquats supportant de longues séries d'évaluation
    - Une capacité d'exécution parallèle et de nouvelles exécutions nécessaires

    - Une échelle appropriée aux exigences d'évaluation
2. **Ressources en personnel** :
    - Des équipes de taille appropriée pour chaque tâche d'évaluation
    - Un support d'ingénierie pour l'implémentation technique
3. **Ressources en temps** :
    - Un temps suffisant pour concevoir, déboguer, exécuter et analyser rigoureusement les évaluations
    - Une allocation de temps proportionnelle à :
        - L'ampleur du risque systémique évalué
        - La complexité de la méthode d'évaluation
        - La stabilité et les caractéristiques de performance du modèle
4. **Support d'ingénierie** :
    - Assistance technique pour les défis d'implémentation
    - Support pour l'inspection des résultats d'évaluation afin d'identifier les bogues potentiels ou les refus du modèle
    - Ressources pour corriger les problèmes qui pourraient conduire à des estimations artificiellement réduites des capacités

Transparence sur l'apport externe dans la prise de décision

Nous consulterons des acteurs externes[^12] à divers moments dans la prise de décisions concernant le développement, le déploiement sur le marché ou l'utilisation de nos modèles d'IA, notamment :

- **Pré-expansion** : Avant d'étendre un modèle d'IA de pointe, nous fournirons un dossier de sécurité pré-expansion à [acteur externe]. Le dossier de sécurité pré-expansion fournira nos preuves, basées sur les évaluations de modèle et les lois de mise à l'échelle, expliquant pourquoi nous pensons qu'une certaine quantité de mise à l'échelle (mesurée en calcul, temps calendaire ou autres quantités) ne franchira pas les seuils de risque pour lesquels nous ne disposons pas encore de mesures de sécurité adéquates. [Acteur externe] aura ... jours pour examiner le dossier de sécurité pré-expansion et fournir un retour non contraignant. Nous veillerons à ce qu'avant toute expansion, nous ayons eu un dossier de sécurité pré-expansion qui démontre que la poursuite de la mise à l'échelle est sûre.
- **Avant le déploiement sur le marché** : Avant de déployer un modèle d'IA de pointe sur le marché, nous fournirons un dossier de sécurité pré-déploiement à [acteur externe]. Le dossier de sécurité pré-déploiement fournira des preuves expliquant pourquoi notre modèle est sûr à déployer en externe et ne posera pas de risque systémique inacceptable. [Acteur externe] aura ... jours pour examiner le dossier de sécurité pré-déploiement et fournir un retour non contraignant.
- **Utilisation** : Avant de déployer en interne un modèle d'IA de pointe pour une utilisation large en recherche, nous fournirons un dossier de sécurité d'utilisation interne pour examen par [acteur externe].

Alternative : À ce stade, nous ne nous engageons pas à recevoir des contributions d'acteurs externes, y compris d'acteurs gouvernementaux, dans la prise de décisions concernant le développement, le déploiement sur le marché ou l'utilisation de nos modèles d'IA.

Nos prévisions

Prévisions actuelles

Articles pertinents :

- Evaluating Frontier Models for Dangerous Capabilities § Expert forecasts of dangerous capabilities
- Measuring AI Ability to Complete Long Tasks
- Mapping AI Benchmark Data to Quantitative Risk Estimates Through Expert Elicitation

Cyberoffensive

Estimation du calendrier : Nous estimons une probabilité de ...% de développer des modèles avec des capacités cyberoffensives suffisantes pour poser le niveau inacceptable dans ... mois, avec la distribution suivante :

- 10e percentile : ...
- 50e percentile : ...
- 90e percentile : ...

Justification : ...

Hypothèses sous-jacentes : ...

Facteurs d'incertitude : ...

Risque CBRN

Estimation du calendrier : Nous estimons avec une confiance de ...% que des modèles auront des capacités suffisantes pour poser le niveau inacceptable pour le CBRN émergeront dans ... mois, avec la distribution suivante :

- 10e percentile : ...
- 50e percentile : ...
- 90e percentile : ...

Justification : ...

Hypothèses sous-jacentes : ...

Facteurs d'incertitude : ...

Perte de contrôle

Estimation du calendrier : Nous évaluons avec une confiance de ...% que des modèles auront des capacités suffisantes pour poser le niveau inacceptable dans ... mois, avec la distribution suivante :

- 10e percentile : ...
- 50e percentile : ...
- 90e percentile : ...

Hypothèses sous-jacentes : ...

Justification : ...

Facteurs d'incertitude : ...

Manipulation nuisible

Estimation du calendrier : Nous prévoyons avec une confiance de ...% que des modèles auront des capacités suffisantes pour poser le niveau inacceptable dans ... mois, avec la distribution suivante :

- 10e percentile : ...
- 50e percentile : ...
- 90e percentile : ...

Hypothèses sous-jacentes : ...

Justification : ...

Facteurs d'incertitude : ...

Travaux futurs

Nous viserons à intégrer les évaluations de modèles pour nos niveaux de risque systémique les plus élevés avec les efforts de prévision, notamment :

Expériences de loi d'échelle

Nous mènerons des expériences systématiques de mise à l'échelle pour comprendre les tendances de capacité à travers :

- Paramètres de taille du modèle
- Allocation de calcul d'entraînement
- Exigences de calcul d'inférence

Ces expériences documenteront les relations observées entre la mise à l'échelle des ressources et l'émergence des capacités, soutenant une prédiction plus précise des capacités futures et des risques.

Comparaisons de générations de modèles

Nous mettrons en œuvre des comparaisons structurées entre les générations consécutives de modèles pour :

- Suivre les trajectoires de croissance des capacités
- Identifier l'accélération ou la décélération dans le développement des capacités
- Documenter les capacités émergentes non présentes dans les générations précédentes

Techniques de prévision supplémentaires

Le cas échéant, nous emploierons d'autres techniques pour estimer quand des capacités, des propensions, des affordances et des effets pourraient émerger dans les modèles futurs, telles que :

- Analyse des tendances dans l'ensemble du domaine de l'IA
- Extrapolation statistique à partir des modèles de capacité observés
- Élicitation structurée d'experts pour la prévision des capacités
- Comparaison systématique avec des benchmarks et des évaluations de tiers

Toutes les techniques de prévision seront mises en œuvre conformément aux méthodes de pointe pertinentes et documentées pour soutenir l'affinement continu de nos approches prédictives.

Recherche exploratoire et red-teaming ouvert

Évaluation exploratoire des modèles

Reconnaissant la nature évolutive de l'évaluation des risques systémiques, nous menons des recherches exploratoires qui vont au-delà de l'évaluation des capacités et des risques connus. Les domaines que nous explorerons comprennent :

1. Tests exploratoires pour des capacités non anticipées ou des comportements émergents

2. Développement de nouvelles méthodologies d'évaluation
3. Investigation de nouvelles voies potentielles de risque systémique
4. Recherche de techniques améliorées pour l'élicitation de modèles
5. Réalisation de méta-recherches sur l'efficacité des méthodes d'évaluation
6. Soutien à la recherche de prévision pour anticiper les capacités futures

Red-teaming ouvert

Nous nous engageons dans des tests adversariaux structurés par le biais de red-teaming ouvert qui :

1. Impliquent divers représentants experts et profanes de :
    - La société civile
    - Le milieu académique
    - La recherche en sécurité
    - Les communautés affectées
2. Emploient des méthodologies conçues pour découvrir :
    - De nouveaux vecteurs d'attaque
    - Des modes de défaillance non anticipés
    - Des comportements émergents
    - Des limites de capacité
3. Fournissent une liberté appropriée pour l'exploration créative de :
    - Scénarios potentiels d'utilisation abusive
    - Vulnérabilités du système
    - Contournement des limitations de sécurité
    - Nouveaux vecteurs de risque

Lorsque les parties prenantes directement affectées par notre modèle d'IA de pointe ne sont pas disponibles pour le red-teaming, nous identifions et engageons des représentants appropriés pour évaluer les impacts potentiels de leur perspective.

Intégration des résultats

Les résultats de la recherche exploratoire et du red-teaming ouvert sont :

1. Documentés et analysés pour leurs implications en matière de risque systémique
2. Incorporés dans nos évaluations de risque systémique
3. Utilisés pour améliorer les méthodologies d'évaluation futures
4. Partagés avec les autorités pertinentes et d'autres parties prenantes selon le cas
5. Appliqués pour améliorer nos stratégies d'atténuation des risques

Partage d'outils et de meilleures pratiques

Notre organisation reconnaît que les risques systémiques sont influencés par les interactions entre plusieurs modèles et systèmes d'IA, y compris les réactions en chaîne potentielles d'incidents ou de dysfonctionnements. L'évaluation et l'atténuation efficaces des risques systémiques nécessitent une collaboration au sein de l'écosystème plus large de l'IA.

Notre approche de partage

Nous partagerons des outils d'évaluation de modèles de pointe et des meilleures pratiques avec les acteurs pertinents de l'écosystème de l'IA, en particulier ceux engagés dans l'évaluation des risques ou l'atténuation des modèles ou systèmes d'IA qui peuvent interagir avec nos modèles d'IA de pointe. Ceux-ci comprennent :

- D'autres fournisseurs de modèles (en particulier les PME)
- Les fournisseurs en aval
- Les évaluateurs externes indépendants
- Les institutions académiques

Notre programme de partage inclura des méthodologies et des procédures établies tout en maintenant notre capacité à effectuer une évaluation et une atténuation rigoureuses des risques pour nos propres modèles. Nous dédierons des ressources d'ingénierie et de support pour faciliter ce partage sans compromettre l'efficacité de notre équipe interne d'évaluation des risques systémiques.

Mécanismes de partage

Jusqu'à ce qu'il y ait des processus établis par l'industrie ou le gouvernement, nous mettrons en œuvre les mécanismes suivants :

1. **Partage ouvert** : Dans la mesure du possible, nous publierons publiquement le code source, la documentation et les matériels de formation pour maximiser l'accessibilité pour tous les acteurs pertinents.
2. **Partage sécurisé** : Pour les outils ou méthodologies plus sensibles qui nécessitent un accès restreint, nous établirons des canaux sécurisés, par exemple par le biais d'organisations industrielles, pour faciliter le partage avec des destinataires qualifiés.

Amélioration continue

Nous examinerons et améliorerons régulièrement notre programme de partage en fonction de :

- Retours des destinataires
- Avancées dans les méthodes d'évaluation

- Évolution des meilleures pratiques
- Conseils des autorités pertinentes

Cette approche nous permet de contribuer de manière significative à l'écosystème plus large de la sécurité de l'IA tout en maintenant les garde-fous nécessaires autour des informations sensibles.

Surveillance post-commercialisation

Nous effectuerons une surveillance post-commercialisation pour recueillir des informations sur les capacités, les propensions, les affordances et les effets du modèle pour évaluer et atténuer les risques systémiques. Notre surveillance :

1. S'assurera que nos modèles ne posent pas de risque systémique inacceptable tel que défini par nos critères d'acceptation des risques
2. Soutiendra le travail de prévision

Méthodes de surveillance

Notre surveillance post-commercialisation utilisera des méthodes appropriées à notre stratégie d'intégration, de publication et de distribution :

[liste selon les besoins, peut-être en incluant certains des éléments ci-dessous]

- **Collecte de retours des utilisateurs** par le biais de canaux structurés et d'enquêtes
- **Canaux de signalement** pour que les utilisateurs signalent des problèmes potentiels
- **Formulaires de rapport d'incident** pour une documentation systématique
- **Programmes de prime aux bogues** pour inciter à la découverte de risques potentiels
- **Évaluations communautaires** et surveillance des tableaux de classement publics
- **Suivi de l'utilisation dans le monde réel** comprenant :
    - Surveillance de l'utilisation du modèle dans les dépôts de logiciels
    - Identification de l'utilisation dans des logiciels malveillants connus
    - Suivi des nouveaux modèles d'utilisation dans les forums publics et les médias sociaux
- **Collaboration académique** avec des chercheurs étudiant les effets de nos modèles
- **Dialogues avec les parties prenantes** avec les groupes affectés
- **Surveillance technique** lorsque c'est faisable, comprenant :
    - Développement de techniques de surveillance préservant la confidentialité
    - Mise en œuvre de filigranes et d'empreintes digitales
    - Analyse des métadonnées lorsque c'est techniquement et légalement approprié

Domaines d'intérêt

Nous recueillerons spécifiquement des informations sur :

1. Les violations des restrictions d'utilisation et les incidents qui en découlent
2. Pour les modèles à code fermé, les aspects pertinents pour l'évaluation des risques qui ne sont pas transparents pour les tiers

Surveillance interne

Pour les systèmes d'IA incorporant nos modèles que nous fournissons ou déployons nous-mêmes, nous surveillerons le modèle dans le cadre de ces systèmes pour évaluer et atténuer efficacement les risques systémiques.

Informations de tiers

Lorsque nos méthodes de surveillance choisies ne fournissent pas suffisamment d'informations, nous chercherons la coopération avec les licenciés, les fournisseurs en aval et les utilisateurs finaux pour recevoir des informations pertinentes, toujours conformément aux lois applicables sur la protection des données. Pour les utilisateurs finaux consommateurs, nous mettrons en œuvre des mécanismes de partage opt-in pour les informations de surveillance pertinentes.

Documentation

Rapports de Sécurité et de Sûreté des Modèles

Nos Rapports de Modèles détailleront l'évaluation des risques et les mesures d'atténuation que nous avons menées pour chaque modèle d'IA de pointe posant potentiellement un risque systémique.

Niveau de détail

Nos Rapports de Modèles fourniront un niveau de détail qui :

1. Est proportionné au niveau de risque systémique que nos modèles atteignent ou sont raisonnablement susceptibles d'atteindre tout au long de leur cycle de vie
2. Permet une compréhension claire de la façon dont nous mettons en œuvre des mesures d'évaluation et d'atténuation des risques systémiques

Chaque Rapport de Modèle contiendra un raisonnement suffisant pour justifier ces déterminations, y compris des explications lorsqu'un niveau de détail inférieur est approprié (comme pour les modèles dérivés en toute sécurité).

Documentation des décisions de publication

Chaque Rapport de Modèle fournira un raisonnement clair et des informations justifiant notre décision de publier un modèle, notamment :

1. Un raisonnement complet démontrant que le risque systémique est acceptable selon nos critères d'acceptation, y compris une explication de la marge de sécurité incorporée
2. Une documentation claire des conditions dans lesquelles notre raisonnement ne tiendrait plus
3. Des informations sur l'implication des évaluateurs externes indépendants dans le processus de prise de décision, y compris comment leurs évaluations ont été incorporées

Documentation de l'évaluation et de l'atténuation des risques systémiques

Nos Rapports de Modèles documenteront en détail :

1. Notre processus de sélection des risques systémiques, y compris la justification de la sélection de risques spécifiques
2. Le processus d'estimation des risques systémiques, justifiant toutes les décisions prises pendant l'évaluation, en particulier concernant les approches qualitatives vs quantitatives
3. Pour les modèles dérivés en toute sécurité, une justification complète de la façon dont les critères pour les modèles d'origine sûrs et les modèles dérivés en toute sécurité sont remplis
4. Les limitations et incertitudes dans notre évaluation des risques, notamment :
    - Les marges de sécurité mises en œuvre
    - Tout niveau inférieur de rigueur dans l'évaluation avec justification
    - Les conséquences pour l'évaluation du modèle
5. Les efforts d'élicitation de modèle pour chaque évaluation
6. Comment les informations sur les systèmes d'IA ont été incorporées pour chaque évaluation
7. Les qualifications des équipes internes et des évaluateurs externes, y compris les niveaux d'accès et les ressources fournies
8. La comparaison des niveaux de risque avec et sans mesures d'atténuation
9. Toutes les mesures d'atténuation mises en œuvre et leurs limitations

De plus, nous documenterons :

10. La base pour conclure que nos évaluations sont à la pointe de la technologie
11. La validité interne, la validité externe et la reproductibilité de chaque évaluation
12. La couverture des contextes d'utilisation et des modalités attendus
13. Les résultats de la recherche exploratoire et du red-teaming ouvert
14. Les outils, les meilleures pratiques et les données partagés avec d'autres pour les évaluations de modèles

Rapports externes

Nos Rapports de Modèles incluront :

1. Les rapports disponibles des évaluateurs externes indépendants qui ont examiné nos modèles avant leur mise sur le marché et des revues de sécurité
2. Lorsqu'aucun évaluateur externe n'a été impliqué, une explication claire de la raison pour laquelle les critères nécessitant une telle implication n'ont pas été remplis, ou pourquoi aucun évaluateur approprié n'a été trouvé
3. Lorsque des évaluateurs externes ont été impliqués, une justification de leur sélection basée sur des critères de qualification, à moins qu'ils n'aient été reconnus par les autorités de régulation

Améliorations algorithmiques

Chaque Rapport de Modèle contiendra des informations de haut niveau sur les améliorations algorithmiques ou autres spécifiques au modèle par rapport à d'autres modèles disponibles sur le marché, lorsque ces informations sont pertinentes pour comprendre les changements significatifs dans le paysage des risques systémiques.

Spécification du comportement prévu du modèle

Nos Rapports de Modèles spécifieront clairement comment nous prévoyons que le modèle fonctionne, notamment :

1. Les principes que le modèle est censé suivre
2. Comment le modèle priorise différents types d'instructions
3. Les sujets sur lesquels le modèle est censé refuser les instructions

Mises à jour des Rapports de Modèles

Nous mettrons à jour les Rapports de Modèles lorsque nous aurons des raisons de croire qu'il y a eu un changement matériel dans le paysage des risques systémiques qui remet en cause notre évaluation initiale, notamment :

1. Des changements significatifs dans les capacités du modèle par le biais de post-entraînement, d'élicitation ou de changements d'affordances
2. Une dérive des données, des améliorations dans les méthodes d'évaluation ou des facteurs suggérant que les évaluations précédentes ne sont plus précises
3. Des améliorations dans les mesures d'atténuation
4. Des incidents graves
5. Des informations de la surveillance post-commercialisation indiquant des changements dans les modèles d'utilisation
6. Des informations provenant de l'utilisation interne du modèle

Chaque Rapport de Modèle mis à jour :

1. Inclura le contenu du rapport précédent, mis à jour lorsque c'est pertinent
2. Inclura de nouveaux rapports des évaluateurs externes indépendants
3. Évaluera si notre Cadre a été correctement respecté
4. Inclura un journal des modifications détaillé avec numéro de version et date
5. Expliquera pourquoi, le cas échéant, nous concluons qu'il n'y a pas eu de changement matériel dans le paysage des risques systémiques

Évaluations d'adéquation

Processus d'évaluation d'adéquation

Lors de la conduite des évaluations d'adéquation, nous prendrons en compte :

1. Les meilleures pratiques dans l'industrie
2. La recherche pertinente et la science de pointe
3. Les incidents ou dysfonctionnements et les rapports d'incidents graves
4. L'expertise externe indépendante disponible

Les évaluations d'adéquation complétées seront fournies à notre organe de direction ou à un autre organe indépendant approprié dans les 5 jours ouvrables suivant leur achèvement.

Évaluation d'adéquation spécifique au modèle

Pour chaque modèle répondant aux conditions de classification (sauf les modèles dérivés en toute sécurité), nous documenterons :

1. Description générale des techniques et des actifs utilisés dans le développement
2. Résultats de l'identification des risques systémiques
3. Prévisions disponibles montrant les capacités, les indicateurs de risque systémique et les niveaux de risque que le modèle est susceptible d'atteindre
4. Résultats disponibles d'analyse des risques systémiques
5. Aperçu de l'analyse des risques systémiques prévue, de la détermination de l'acceptation et des mesures d'atténuation techniques
6. Caractérisation des risques systémiques qui peuvent survenir pendant le développement, avec évaluation des risques significatifs pendant la phase de développement

Nous mettrons à jour ces évaluations lorsque nous atteindrons des jalons définis si :

1. Des risques systémiques significatifs surviennent pendant le développement
2. Il y a eu des changements matériels dans les informations ci-dessus
3. Aucune évaluation n'a été complétée au cours des 12 dernières semaines
4. Le modèle n'a pas été mis à disposition sur le marché

Les mises à jour se concentreront sur les risques en phase de développement, en particulier l'automatisation de la recherche en IA et les mesures de protection de sécurité.

Évaluation d'adéquation du Cadre

Nous mènerons des évaluations d'adéquation du Cadre :

1. Lorsque nous serons informés de changements matériels qui pourraient compromettre l'adéquation du Cadre
2. Tous les 12 mois à partir de notre première mise sur le marché d'un modèle

Ces évaluations évalueront :

1. Si le Cadre traite tous les composants requis et est cohérent avec les exigences réglementaires
2. Si le Cadre reste proportionné aux risques des modèles que nous prévoyons de publier dans les 12 prochains mois
3. S'il y a une forte raison de croire que le Cadre sera respecté

Conservation

Nous maintiendrons une documentation complète comprenant :

1. La documentation requise dans le cadre d'autres cadres réglementaires
2. Notre Cadre, y compris les versions précédentes
3. Les Rapports de Modèles, y compris les versions précédentes
4. Toutes les évaluations d'adéquation complétées

Toute la documentation sera conservée pendant au moins 12 mois après le retrait du modèle.

Attribution des responsabilités en matière de risque systémique

Définition des responsabilités

Nous définirons clairement les responsabilités pour la gestion des risques systémiques à tous les niveaux de l'organisation :

1. **Surveillance des risques** : Le [comité spécifique] de notre organe de direction supervisera les activités de gestion des risques systémiques

2. **Propriété des risques** : [Rôles de gestion spécifiques] assumeront la responsabilité directe de la gestion des risques systémiques
3. **Support et surveillance** : [Rôle spécifique] soutiendra et surveillera la gestion des risques systémiques, soutenu par une fonction de risque centrale
4. **Assurance** : [Rôle spécifique] fournira une assurance sur l'adéquation de la gestion des risques systémiques, soutenu par une fonction d'audit indépendante

Allocation de ressources appropriées

Notre direction supervisera l'allocation de ressources appropriées proportionnelles aux niveaux de risque systémique :

1. Ressources humaines avec l'expertise pertinente
2. Ressources financières
3. Accès aux informations et connaissances nécessaires
4. Ressources informatiques pour une évaluation approfondie

Promotion d'une culture saine du risque

Nous promouvrons une approche équilibrée du risque systémique, encourageant une sensibilisation appropriée au risque sans recherche excessive ou aversion excessive au risque :

1. Définir le ton depuis la direction concernant la gestion équilibrée des risques
2. Permettre une communication efficace et la remise en question des décisions de risque
3. Créer des incitations décourageant la prise de risque excessive
4. Sonder régulièrement le personnel sur la sensibilisation aux risques et le confort à soulever des préoccupations
5. Maintenir des canaux de signalement actifs avec un suivi approprié

[ajouter des spécificités d'entreprise selon les besoins]

Signalement des incidents graves

Méthodes d'identification des incidents graves

Nous mettrons en œuvre des méthodes appropriées à notre modèle d'affaires pour suivre les informations sur les incidents graves :

1. Traçage des sorties du modèle à l'aide de filigranes, de métadonnées ou d'autres techniques de provenance

2. Conduite d'une journalisation préservant la confidentialité et d'une analyse des métadonnées lorsque c'est possible
3. Examen des sources d'information externes, y compris les rapports de police et de médias, les médias sociaux et les bases de données d'incidents
4. Facilitation du signalement par les fournisseurs en aval, les utilisateurs et les tiers

Informations pertinentes pour le suivi, la documentation et le signalement des incidents graves

Nous suivrons, documenterons et signalerons au moins :

1. Les dates de début et de fin des incidents
2. Les dommages résultants et les groupes affectés
3. La chaîne d'événements menant à l'incident
4. La version du modèle impliquée
5. La description du matériel montrant l'implication de notre modèle
6. Nos actions de réponse et plans
7. Les recommandations pour la réponse réglementaire
8. L'analyse des causes profondes, y compris les sorties, les entrées et les défaillances des mesures de protection
9. Les quasi-accidents connus

Nous enquêterons sur les causes et les effets des incidents pour informer l'analyse future des risques, avec un niveau de détail proportionnel à la gravité de l'incident.

Délais de signalement

Pour les incidents graves, lorsque convenu avec les autorités pertinentes, nous soumettrons des rapports initiaux contenant les points d'information 1-7 ci-dessus dès que possible, et au plus tard :

1. Dans les 2 jours pour une perturbation grave d'une infrastructure critique
2. Dans les 10 jours pour un préjudice grave à l'intégrité physique ou une relation causale suspectée
3. Dans les 15 jours pour un préjudice grave à la santé, des violations des droits fondamentaux, ou des dommages graves à la propriété/environnementaux
4. Dans les 5 jours pour des incidents graves de cybersécurité ou l'exfiltration des poids du modèle

Nous soumettrons des rapports intermédiaires toutes les 4 semaines jusqu'à la résolution, et des rapports finaux dans les 60 jours suivant la résolution, couvrant toutes les exigences d'information.

Période de conservation

Nous conserverons toute la documentation produite dans le cadre de cet engagement pendant au moins 36 mois à compter de la date de documentation ou de la date de l'incident, selon la plus tardive.

Protections contre les représailles

Nous ne prendrons pas de mesures de représailles contre tout travailleur fournissant des informations sur les risques systémiques aux autorités de régulation. Nous informerons annuellement les travailleurs des canaux réglementaires désignés pour recevoir ces informations.

Notifications

Nous partagerons des informations sur les modèles posant potentiellement des risques systémiques avec les autorités comme requis par les règles applicables, telles que le Code de Pratique pour l'IA à Usage Général, section *Notifications*.

Transparence publique

Nous publierons des informations sur les risques systémiques qui améliorent :

1. La sensibilisation et la connaissance publiques des risques
2. La résilience sociétale contre les risques
3. La détection des risques avant et lorsqu'ils se matérialisent

Plus précisément, nous publierons sur notre site web :

1. Les versions nouvelles ou mises à jour de notre Cadre
2. Les Rapports de Modèles avec des rédactions si nécessaire pour les informations qui pourraient augmenter le risque, y compris des informations suffisantes pour la compréhension publique de :
    - Comment le risque évalué du modèle se rapporte à nos niveaux de risque.
    - Pour les niveaux de risque non atteints, une explication de pourquoi les preuves recueillies montrent que le niveau de risque n'est pas atteint.
    - La qualité de l'élicitation.
    - Tous les rapports de tiers.

Les informations publiées incluront des rédactions nécessaires pour empêcher l'augmentation du risque de compromettre les mesures de sécurité ou de sûreté ou de révéler des informations commerciales sensibles disproportionnées par rapport au bénéfice sociétal.

Amélioration et mise à jour du Cadre

Au fil du temps, nous mettrons à jour et améliorerons notre Cadre avec l'intention de le rendre plus efficace pour évaluer et atténuer les risques systémiques en :

1. Incorporant les enseignements tirés de l'application du Cadre à nos modèles d'IA de pointe
2. Affinant et validant les prévisions au fil du temps en les comparant aux tendances historiques et à d'autres estimations
3. Mettant régulièrement à jour le Cadre pour incorporer des mesures et des procédures de pointe pour l'évaluation et l'atténuation des risques

Une version mise à jour du Cadre sera considérée comme complète lorsque :

1. Elle aura été examinée par notre équipe de sécurité et de risque d'IA
2. Elle aura été approuvée par notre organe de direction agissant dans sa fonction de supervision
3. Tous les changements auront été documentés avec des justifications
4. Le journal des modifications aura été mis à jour avec un numéro de version et une date de changement

Sélection des risques systémiques

Lors de la planification du développement d'un nouveau modèle avec le potentiel de poser un risque systémique, nous identifierons les risques systémiques potentiels spécifiques aux capacités à fort impact de notre modèle d'IA de pointe en :

1. Examinant une gamme de risques potentiels, tels que ceux décrits dans les Annexes 1.1 et 1.2 du Code de Pratique pour l'IA à Usage Général
2. Analysant les informations sur les risques présentés par des modèles similaires sur le marché
3. Consultant la recherche actuelle et les opinions d'experts sur les risques émergents
4. Engageant les parties prenantes susceptibles d'être affectées par notre modèle d'IA de pointe

Détermination des scénarios de risque systémique

Pour chaque risque systémique identifié, nous développerons des scénarios de risque détaillés qui :

1. Caractérisent le type, la nature et les sources du risque
2. Décrivent les voies vers des préjudices, y compris les mauvais usages par négligence et intentionnels raisonnablement prévisibles
3. Documentent comment le développement, la mise à disposition sur le marché et/ou l'utilisation de notre modèle d'IA de pointe pourraient produire le risque systémique
4. Identifient les capacités du modèle, les propensions, les affordances et les facteurs contextuels qui détermineraient si le risque est présent

5. Identifient quelles mesures d'atténuation conduiraient le plus efficacement à un risque acceptable

Notre méthodologie de modélisation des risques incorpore des approches tant quantitatives que qualitatives, combinant la modélisation des menaces avec des évaluations de modèle spécifiques pour créer des scénarios de risque complets qui informent nos stratégies d'analyse et d'atténuation subséquentes.

Journal des modifications

Version	Date	Changements	Justification
1.0	2025-04-07	Cadre initial	…

[^1]: Le Cadre s'applique à tous les modèles que nous développons qui, une fois pleinement entraînés, pourraient posséder des capacités générales comparables à celles des modèles de pointe des 5 à 15 principales entreprises d'IA.

[^2]: LLM Cyber Evaluations Don't Capture Real-World Risk

[^3]: Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models

[^4]: Navigating Dual-Use Refusal Policy for AI Systems in Cybersecurity

[^5]: Adversaries Can Misuse Combinations of Safe Models

[^6]: Dirty bomb

[^7]: Adversaries Can Misuse Combinations of Safe Models

[^8]: Wired couvre un exemple d'un stagiaire humain qui a tenté de saboter l'entraînement d'un modèle d'IA et a ensuite remporté le Prix du Meilleur Article à NeurIPS.

[^9]: Voir également, Do the Returns to Software R&D Point Towards a Singularity?

[^10]: A sketch of an AI control safety case

[^11]: Notez également que permettre aux utilisateurs externes d'utiliser les capacités de R&D en IA, même de manière extensive, peut ne pas être indésirable et peut être utile pour traiter les risques liés à la concentration du pouvoir.

[^12]: Les acteurs externes pertinents à considérer pourraient inclure des membres du Réseau International des Instituts de Sécurité de l'IA (Australie, Canada, Chine, Union européenne, France, Japon, Kenya, République de Corée, Singapour, Royaume-Uni et États-Unis, qui ont différents niveaux

d'affiliation gouvernementale), des collectifs industriels tels que le Forum des Modèles de Pointe, et des organisations de la société civile telles qu'Apollo Research et METR.