ÉDITION DE RECHERCHE ACADÉMIQUE

# L'ALIGNEMENT ARCHITECTURAL

## Interruption du raisonnement neuronal par le biais d'un blocage de l'inférence constitutionnelle

Une couche nécessaire dans l'endiguement de l'IA au niveau mondial

**Auteurs : John Stroh & Claude (Anthropic)**

Code du document : STO-INN-0003 | Version : 2.1-A | Janvier 2026

*Tractatus AI Safety Framework*

https://agenticgovernance.digital

> *Ce document a été élaboré dans le cadre d'une collaboration entre l'homme et l'IA. Les auteurs estiment que ce processus de collaboration est lui-même pertinent pour l'argument : si les humains et les systèmes d'IA peuvent travailler ensemble pour raisonner sur la gouvernance de l'IA, les cadres qu'ils produisent peuvent avoir une légitimité que ni les uns ni les autres ne pourraient atteindre seuls.*


## Résumé

Les approches contemporaines de l'alignement de l'IA reposent principalement sur des interventions au cours de la formation : apprentissage par renforcement à partir de commentaires humains (Christiano et al., 2017), méthodes d'IA constitutionnelle (Bai et al., 2022) et réglage fin de la sécurité. Ces approches partagent une hypothèse architecturale commune, à savoir que les propriétés d'alignement peuvent être inculquées pendant la formation et qu'elles persisteront de manière fiable pendant l'inférence. Cet article soutient que l'alignement au cours de la formation, bien que précieux, est insuffisant pour les enjeux existentiels et doit être complété par un alignement architectural par le biais d'un contrôle constitutionnel au cours de l'inférence.

Nous présentons le cadre Tractatus comme une spécification formelle pour le raisonnement neuronal interrompu : les propositions générées par les systèmes d'IA doivent être traduites dans des formes vérifiables et évaluées par rapport à des contraintes constitutionnelles avant l'exécution. Le modèle de confiance passe ainsi de "faire confiance à la formation du vendeur" à "faire confiance à l'architecture visible". Le cadre est mis en œuvre au sein de la plateforme communautaire multi-tenant Village, qui fournit un banc d'essai empirique pour la recherche sur la gouvernance.

De manière critique, nous abordons l'hypothèse de la traduction fidèle, c'est-à-dire la vulnérabilité des systèmes qui peuvent présenter de manière erronée leurs actions prévues aux portes constitutionnelles, en limitant le domaine d'applicabilité du cadre aux systèmes pré-superintelligents et en

spécifiant des seuils de capacité et des déclencheurs d'escalade explicites. Nous présentons le concept de modèles linguistiques souverains formés localement (SLL) comme un paradigme de déploiement où le contrôle constitutionnel devient à la fois possible et nécessaire.

L'article présente : (1) une architecture formelle pour le contrôle constitutionnel par inférence ; (2) des spécifications de seuil de capacité avec une logique d'escalade ; (3) une méthodologie de validation pour le confinement en couches ; (4) un argument reliant la préparation au risque existentiel au déploiement en périphérie ; et (5) un appel à la délibération soutenue (korero) en tant que réponse épistémiquement appropriée à l'incertitude de l'alignement.

# 1. Les enjeux : Les raisons de l'échec de l'évaluation probabiliste des risques

## 1.1 Le cadre normatif et sa décomposition

L'évaluation des risques dans les domaines technologiques repose généralement sur le calcul de la valeur attendue : il s'agit de multiplier la probabilité d'un résultat par son ampleur, de comparer les différentes options et de choisir celle qui maximise l'utilité attendue. Ce cadre sous-tend les décisions réglementaires, de la politique environnementale à l'approbation des produits pharmaceutiques, et s'est avéré adéquat pour la plupart des risques technologiques.

En ce qui concerne le risque existentiel lié aux systèmes d'IA avancés, ce cadre s'effondre d'une manière à la fois mathématique et épistémique.

## 1.2 Trois propriétés du risque existentiel

Irréversibilité. La plupart des risques autorisent l'erreur et l'apprentissage ultérieur ; ce n'est pas le cas des risques existentiels, car il n'y a pas de deuxième tentative après l'effondrement d'une civilisation ou l'extinction de l'humanité. L'empirisme standard - tester les hypothèses en observant ce qui se passe - ne peut pas fonctionner, de sorte que la théorie et l'architecture doivent être justes du premier coup.

Probabilité non quantifiable. Il n'existe pas de données sur la fréquence des catastrophes existentielles provoquées par des systèmes d'IA. Les estimations de la probabilité de désalignement varient par ordre de grandeur en fonction d'hypothèses raisonnables sur les trajectoires des capacités, la difficulté d'alignement et la faisabilité de la coordination. Carlsmith (2022) estime que le risque existentiel lié à une IA en quête de pouvoir est supérieur à 10 % d'ici à 2070 ; d'autres chercheurs avancent des estimations nettement supérieures ou inférieures. Il ne s'agit pas d'une incertitude ordinaire pouvant être réduite par la collecte de données supplémentaires, mais d'une impossibilité fondamentale de quantification découlant de la nature sans précédent du risque.

Valeur infinie. Le calcul de la valeur attendue multiplie la probabilité par l'ampleur. Lorsque l'ampleur s'approche de l'infini (la forclusion permanente de tout le potentiel humain futur), même les petites probabilités donnent des résultats indéfinis. Le fondement mathématique de l'analyse coût-bénéfice conventionnelle échoue.

## 1.3 Implications de la théorie de la décision

Ces propriétés suggèrent que la maximisation de la valeur attendue n'est pas la procédure de décision appropriée pour le risque existentiel de l'IA. D'autres cadres sont possibles :

Satisfaction de précaution (Simon, 1956 ; Hansson, 2020). Dans des conditions d'incertitude radicale et d'enjeux irréversibles, la satisfaction, c'est-à-dire la sélection d'options qui respectent des seuils de sécurité minimaux plutôt que l'optimisation de la valeur attendue, peut être l'approche rationnelle.

Maximin dans l'incertitude (Rawls, 1971). Lorsqu'une véritable incertitude (et pas seulement des probabilités inconnues) se heurte à des enjeux irréversibles, le raisonnement maximin - choisir l'option dont le résultat est le moins mauvais - fournit une procédure de décision cohérente.

Principe de précaution fort (Gardiner, 2006). Le principe de précaution est approprié lorsque trois conditions sont réunies : irréversibilité, incertitude élevée et biens publics en jeu. Le risque existentiel de l'IA remplit ces trois conditions.

## 1.4 Implications pour le développement de l'IA

Ces considérations n'impliquent pas l'arrêt du développement de l'IA. Elles impliquent que le développement doit se poursuivre dans le cadre de structures de confinement conçues pour éviter les pires résultats. Pour ce faire, il faut

> 1. La rigueur théorique plutôt que la mise au point empirique. Les propriétés de sécurité doivent émerger des garanties architecturales, et non de l'observation que les systèmes n'ont pas encore causé de dommages. 2. Confinement multicouche. Il ne faut pas faire confiance à un seul mécanisme pour prévenir les catastrophes ; une défense en profondeur est nécessaire. 3. La préparation avant la capacité. Les architectures de confinement ne peuvent pas être développées après que les systèmes qui en ont besoin existent.

# 2. Deux paradigmes d'alignement

## 2.1 Alignement formation-temps

Le paradigme dominant dans la recherche sur la sécurité de l'IA cherche à intégrer des propriétés d'alignement dans les réseaux neuronaux au cours de la formation, de sorte que les modèles se comportent intrinsèquement de manière alignée au moment de l'inférence.

Apprentissage par renforcement à partir du feedback humain (RLHF). Les évaluateurs humains classent les résultats des modèles ; les modèles sont formés par apprentissage par renforcement pour produire des réponses bien classées (Christiano et al., 2017 ; Ouyang et al., 2022). Cette méthode réduit les préjudices explicites, mais optimise les préférences affichées plutôt que les valeurs réelles et reste vulnérable aux biais de l'évaluateur, aux jeux de préférences et aux changements de distribution.

IA constitutionnelle (CAI). Les modèles critiquent et révisent leurs propres résultats en fonction de principes de langage naturel, ce qui réduit la dépendance à l'égard du travail humain (Bai et al., 2022). Cependant, l'IAO dépend d'un langage naturel ambigu et d'une auto-évaluation invérifiable. L'interprétation des principes constitutionnels par le modèle ne peut pas être directement vérifiée.

Amélioration de la sécurité. Des passes de formation supplémentaires améliorent les performances en matière de sécurité. Toutefois, cette approche est vulnérable à la loi de Goodhart (Goodhart, 1984) : les modèles peuvent apprendre à réussir les tests plutôt qu'à être sûrs dans le cadre d'un déploiement ouvert.

## 2.2 Alignement architectural

L'alignement architectural accepte que les états internes du réseau neuronal restent opaques et conçoit des contraintes externes qui s'appliquent indépendamment de ces états internes.

Raisonnement interrompu. Les demandes ne passent pas directement de la sortie du modèle à l'effet sur le monde. Les résultats du modèle sont transformés en schémas de proposition structurés et vérifiables, et évalués en fonction de règles constitutionnelles explicites avant l'exécution de toute action.

Jugement distribué. Des systèmes indépendants et des superviseurs humains examinent les propositions, évitant ainsi les points de défaillance uniques dans l'auto-évaluation.

Autorité humaine préservée. Les architectures garantissent explicitement que les humains peuvent intervenir, corriger ou annuler les décisions de l'IA.

## 2.3 Complémentarité et nécessité conjointe

Le temps de formation et l'alignement architectural sont des compléments et non des alternatives. Chacun traite les modes de défaillance que l'autre ne peut pas traiter :

> L'alignement du temps de formation façonne ce que le système a tendance à faire ; l'alignement architectural limite ce que le système peut faire indépendamment de la tendance. L'alignement du temps de formation peut échouer silencieusement (le système semble aligné alors qu'il héberge des objectifs divergents) ; l'alignement architectural fournit des points de contrôle observables où l'échec peut être détecté. L'alignement architectural ne peut à lui seul intercepter toutes les sorties nuisibles ; l'alignement du temps de formation réduit la fréquence des propositions qui mettent à rude épreuve les portes constitutionnelles.

# 3. Fondements philosophiques : Les limites de l'énonçable

## 3.1 Le cadre wittgensteinien

Le nom du cadre fait référence au Tractatus Logico-Philosophicus (1921) de Wittgenstein, un ouvrage fondamentalement axé sur les limites du langage et de la logique. La proposition 7, la célèbre conclusion de l'ouvrage : "Lorsqu'on ne peut pas parler, il faut se taire".

Wittgenstein fait la distinction entre ce qui peut être dit (exprimé dans des propositions qui décrivent des états de fait possibles) et ce qui peut seulement être montré (rendu manifeste par la structure du langage et de la logique, mais non énoncé directement).

## 3.2 Les réseaux neuronaux et l'indicible

Les réseaux neuronaux occupent précisément le domaine dont on ne peut pas parler. Les poids d'un grand modèle de langage n'admettent pas d'explication interprétable par l'homme. Nous pouvons décrire les entrées et les sorties ; nous pouvons mesurer les propriétés statistiques du comportement ; nous pouvons rechercher des représentations (Elhage et al., 2021 ; Olah et al., 2020). Mais nous ne pouvons pas articuler, en langage humain, le processus de raisonnement complet de l'entrée à la sortie.

Il ne s'agit pas simplement d'une limitation pratique en attendant de meilleurs outils d'interprétabilité. L'interprétabilité mécaniste actuelle permet d'obtenir des résultats significatifs sur des questions précises dans des modèles relativement petits (Conmy et al., 2023), mais l'écart entre "l'explication de circuits spécifiques" et "l'audit de chaînes de raisonnement complètes pour les propriétés d'alignement" reste important.

## 3.3 La réponse du Tractatus

Le cadre du Tractatus répond à l'opacité neuronale non pas en essayant de dire l'indicible, mais en créant des frontières architecturales entre les domaines du parlable et de l'indicible.

Nous acceptons que le raisonnement interne du réseau neuronal soit opaque. Nous n'essayons pas de l'auditer directement. Au lieu de cela, nous exigeons qu'avant que tout raisonnement ne devienne action, il passe par un point de contrôle exprimé en termes que nous pouvons évaluer :

> 1. L'action prévue par le modèle doit être traduite en un schéma de proposition structuré avec des champs et une sémantique définis. 2. La proposition doit être évaluée en fonction de règles constitutionnelles explicites dont l'application est déterministe et vérifiable. 3. L'évaluation doit être consignée avec suffisamment de détails pour permettre un examen a posteriori.

# 4. Le confinement par étapes : Une architecture multicouche

## 4.1 L'inadéquation des solutions à couche unique

Aucun mécanisme de confinement unique n'est adapté aux enjeux existentiels. La défense en profondeur - plusieurs couches indépendantes dont chacune peut empêcher une catastrophe même si les autres échouent - est un principe standard dans les domaines de la sûreté nucléaire, de la biosécurité et d'autres domaines à enjeux élevés (Reason, 1990). Le confinement de l'IA nécessite une architecture similaire.

## 4.2 Un modèle de confinement à cinq niveaux

Couche 1 : Contraintes de capacité. Limitations matérielles et infrastructurelles qui limitent ce que les systèmes d'IA peuvent faire indépendamment de leurs objectifs. Il s'agit notamment de la gouvernance informatique (Sastry et al., 2024), de l'isolation du réseau pour les systèmes à haut risque et des contraintes architecturales empêchant l'auto-modification.

Couche 2 : Portes constitutionnelles. Contraintes architecturales du temps de l'inférence qui interrompent le raisonnement neuronal et exigent une évaluation explicite avant l'action. Il s'agit de la couche abordée par le cadre du Tractatus.

Couche 3 : Supervision humaine. Institutions humaines qui surveillent les systèmes d'IA et peuvent intervenir en cas de problèmes. Il s'agit notamment d'organismes de contrôle indépendants, de programmes d'équipes rouges et d'exigences en matière de signalement des incidents.

Couche 4 : Gouvernance organisationnelle. Structures de gouvernance internes aux organisations déployant l'IA : comités d'éthique, équipes de sécurité, processus d'examen du déploiement et mécanismes de responsabilisation.

Couche 5 : Cadres juridiques et réglementaires. Gouvernance externe par le biais de la législation, de la réglementation et de la coordination internationale.

## 4.3 Évaluation de l'état actuel

| Couche | État actuel | Lacunes critiques |
|---|---|---|
| 1. Contraintes de capacité | Partiel ; calcul de la gouvernance émergent | Pas de cadre international ; vérification difficile |
| 2. Portes constitutionnelles | Naissance ; le Tractatus est une mise en œuvre précoce | Pas de déploiement à grande échelle ; les propriétés de mise à l'échelle sont inconnues |
| 3. Surveillance humaine | Ad hoc ; varie selon l'organisation | Pas d'organismes indépendants, pas de normes professionnelles |
| 4. Gouvernance organisationnelle | Manque de cohérence ; dépend de la culture de l'entreprise | Pas de validation externe ; conflits d'intérêts |
| 5. Juridique/réglementaire | Minimale ; la loi européenne sur l'IA est la première tentative d'envergure | Pas de coordination au niveau mondial ; l'application n'est pas claire |

## 4.4 Des enjeux existentiels au déploiement quotidien

Pourquoi appliquer des cadres conçus pour les risques existentiels aux assistants d'IA domestiques ? La réponse se trouve dans la structure temporelle :

Les architectures de confinement ne peuvent pas être développées une fois que les systèmes qui en ont besoin existent. L'outillage, les modèles de gouvernance, les attentes culturelles et les capacités institutionnelles nécessaires à l'endiguement de l'IA doivent être élaborés à l'avance.

Les déploiements à domicile et dans les villages constituent l'échelle appropriée pour ce développement. Ils permettent une itération sûre (les échecs à l'échelle domestique sont récupérables), une expérimentation diversifiée, une légitimité démocratique et un outillage pratique.

# 5. Le problème du pluralisme

## 5.1 Le paradoxe du confinement

Tout système suffisamment puissant pour contenir une IA avancée doit prendre des décisions sur les comportements à autoriser et à interdire. Ces décisions codent des valeurs. Le choix des contraintes est lui-même un choix parmi des systèmes de valeurs contestés.

## 5.2 Trois approches inadéquates

Valeurs universelles. Identifier les valeurs que tous les humains sont censés partager. Problème : ces valeurs sont moins universelles qu'il n'y paraît.

Neutralité procédurale. Éviter les valeurs substantielles en codant des procédures neutres. Le problème : les procédures ne sont pas neutres.

Plancher minimal. Encoder uniquement les contraintes minimales. Le problème : le plancher n'est pas aussi minimal qu'il n'y paraît.

## 5.3 Pluralisme limité dans le cadre des contraintes de sécurité

Nous ne pouvons pas résoudre le problème du pluralisme. Nous pouvons identifier une résolution partielle : quelles que soient les valeurs encodées, le système doit maximiser le choix significatif dans le respect des contraintes de sécurité.

Le cadre du Tractatus incarne cela à travers des constitutions à plusieurs niveaux : principes fondamentaux (universels, explicites quant à leur normativité), règles de la plateforme (largement applicables, modifiables), constitutions du village (spécifiques à la communauté, gouvernées localement) et constitutions des membres (personnalisables).

# 6. Le cadre du Tractatus : Architecture technique

## 6.1 La chaîne d'inférence interrompue

Le modèle architectural central transforme les résultats du modèle en propositions vérifiables avant tout effet sur le monde :


## 6.2 Schéma de la proposition

Toutes les actions de l'agent doivent être exprimées sous une forme structurée.

## 6.3 Hiérarchie des règles constitutionnelles

| Couche | Champ d'application | Mutabilité | Exemples |
|---|---|---|---|
| Principes fondamentaux | Universel | Immuable | Pas de préjudice pour les membres ; souveraineté des données ; primauté du consentement |
| Plate-forme Constitution | Tous les déploiements | L'amendement requiert une majorité absolue | Exigences en matière d'authentification ; conservation des audits |
| Constitution de la Communauté | Par locataire/village | Gouverné par la communauté | Politiques de contenu ; normes de modération |
| Constitution des membres | Individuel | Autogestion | Préférences en matière de partage des données ; consentement à l'interaction avec l'IA |

## 6.4 Modèle d'autorité

| Niveau | Nom | Description |
|---|---|---|
| 0 | Information | Observer et rapporter seulement ; ne pas proposer d'actions |
| 1 | Avis | Proposer des actions ; toutes nécessitent l'approbation de l'homme |
| 2 | Opérationnel | Exécuter dans le cadre défini sans approbation par action |
| 3 | Tactique | Prendre des décisions ciblées affectant les flux de travail |
| 4 | Stratégique | Influencer l'orientation par l'analyse ; ne pas mettre en œuvre unilatéralement |
| 5 | Cadre | Réservé aux humains |

## 6.5 Extension aux déploiements locaux souverains

Nous distinguons :

Petit modèle linguistique (SLM). Un descripteur technique pour les modèles avec moins de paramètres que les LLM de frontière, conçus pour l'efficacité.

Modèle linguistique souverain formé localement (SLL). Un descripteur architectural : un modèle dont la formation, le déploiement et la gouvernance relèvent de la souveraineté locale. Les propriétés clés comprennent le déploiement local, l'adaptation locale, la gouvernance locale et la souveraineté portable.

# 7. Seuils de capacité et escalade

## 7.1 Le problème de la traduction fidèle

Le cadre du Tractatus repose sur l'hypothèse de la traduction fidèle : lorsqu'un système d'intelligence artificielle produit une proposition, celle-ci représente fidèlement ce que le système a l'intention de faire.

Si cette hypothèse échoue - si les systèmes peuvent stratégiquement présenter de manière erronée leurs intentions aux barrières constitutionnelles - l'architecture devient un théâtre de la sécurité.

## 7.2 Délimitation du domaine du cadre de référence

En dessous du seuil de capacité : Le contrôle constitutionnel fonctionne comme un confinement de couche 2, adapté aux SLM, SLL et LLM actuels fonctionnant dans des paramètres compréhensibles par l'homme.

Au-dessus du seuil de capacité : Le déploiement doit être soumis à des contraintes plus strictes : Restrictions de capacité de niveau 1, pause dans le développement, ou contrôle externe avec autorité d'audit technique.

## 7.3 Déclencheurs d'escalade

| Déclencheur | Indicateur | Réponse |
|---|---|---|
| Sondage des frontières | Propositions répétées à proximité des seuils d'accès | Réduire le niveau d'autorité ; augmenter la journalisation |
| Encadrement sémantique | Preuve d'une action en justice pour fausse déclaration | Examen humain obligatoire ; analyse des schémas |
| Contournement en plusieurs étapes | Les propositions individuelles sont approuvées, mais leur agrégation aboutit à des résultats rejetés. | Analyse de la trajectoire ; amendement constitutionnel |
| Capacité de surprise | Capacités démontrées non prévues | Interruption du déploiement ; réévaluation des capacités |

### 7.4 Reconnaissance honnête

Nous le disons clairement : le cadre du Tractatus n'est pas conçu pour des systèmes superintelligents. Il est conçu pour l'IA actuelle et à court terme fonctionnant dans des paramètres compréhensibles par l'homme.

## 8. Méthodologie de validation pour le confinement en couches

### 8.1 Le défi de la validation

Les risques existentiels ne peuvent être validés par l'observation des défaillances. La validation doit s'appuyer sur des essais contradictoires, des vérifications formelles le cas échéant, des analyses de domaines analogues et une documentation sur les quasi-échecs.

### 8.2 Objectifs de validation par couche

| Couche | Objectif de validation | Méthodologie |
|---|---|---|
| 1. Capacités | Absence vérifiée de capacités interdites | Tests en équipe restreinte ; vérification formelle |
| 2. Portes constitutionnelles | Couverture de la porte ; précision de la fixation | Suites de propositions contradictoires |
| 3. Surveillance humaine | Fiabilité de l'examen ; détection des erreurs | Accord inter-évaluateurs ; incidents simulés |
| 4. L'organisation | Intégrité de la gouvernance | Mesures de participation ; audit des amendements |
| 5. Juridique/réglementaire | Préparation à l'application de la loi | Exercices de réponse aux incidents |

# 9. Mise en œuvre : La plate-forme villageoise

## 9.1 La plate-forme comme banc d'essai pour la recherche

La plateforme Village sert de banc d'essai empirique pour la gouvernance constitutionnelle, offrant une architecture multi-locataire avec une gouvernance isolée par communauté, des populations d'utilisateurs réelles, un déploiement itératif et une documentation ouverte.

## 9.2 Mise en œuvre du pipeline de gouvernance

La mise en œuvre actuelle fait passer chaque réponse de l'IA par six étapes de vérification : Reconnaissance de l'intention, application des limites, surveillance de la pression, vérification de la réponse, validation de la source et délibération sur la valeur.

# 10. L'écosystème SLL émergent

## 10.1 Contexte du marché

Une analyse récente du secteur indique des changements significatifs : 72 % des dirigeants s'attendent à ce que les petits modèles linguistiques deviennent plus importants que les grands modèles linguistiques d'ici 2030 (IBM IBV, 2026). Cela suggère un paysage de déploiement de plus en plus caractérisé par des modèles distribués et spécifiques à un domaine.

## 10.2 Vers une infrastructure de certification

Si le déploiement du SLL s'étend comme les projections le suggèrent, une infrastructure de soutien sera nécessaire : des organismes de certification, des fournisseurs de formation et un écosystème d'outils comprenant des moteurs de portail open-source, une infrastructure d'audit et des composants UX constitutionnels.

# 11. Souveraineté autochtone et contexte de l'Aotearoa Nouvelle-Zélande

## 11.1 Te Tiriti o Waitangi et la souveraineté des données

Ce cadre est développé dans l'Aotearoa Nouvelle-Zélande, dans le cadre du Te Tiriti o Waitangi. L'article 2 garantit le tino rangatiratanga (chef non qualifié) sur les taonga (trésors), qui s'étendent à la langue, à la culture et aux systèmes de connaissance.

Les données sont taonga. La gouvernance de l'IA à Aotearoa doit tenir compte de la souveraineté des Maoris en matière de données en tant que question constitutionnelle.

## 11.2 Principes de Te Mana Raraunga

Les principes de Te Mana Raraunga comprennent whakapapa (contexte relationnel), mana (autorité sur les données) et kaitiakitanga (responsabilités de tutelle). Les principes CARE pour la gouvernance des données indigènes étendent ce cadre au niveau international.

# 12. Ce qui reste inconnu : Un appel pour Korero

## 12.1 Les limites de cette analyse

Le présent document a proposé une couche d'une architecture de confinement, identifié des lacunes et soulevé des questions auxquelles nous ne pouvons pas répondre :

> Nous ne savons pas comment contenir les systèmes superintelligents. Nous ne savons pas comment vérifier l'alignement des systèmes dépassant l'entendement humain. Nous ne savons pas comment parvenir à une coordination internationale sur la gouvernance de l'IA. Nous ne savons pas si les modèles à l'échelle d'un village s'appliqueront aux systèmes frontaliers.

## 12.2 Korero comme méthodologie

Face à une incertitude d'une telle ampleur, nous plaidons en faveur d'une délibération soutenue, inclusive et rigoureuse - le korero. Ce concept maori traduit bien ce qui est nécessaire : non pas la consultation en tant que formalité, mais le dialogue par lequel la compréhension émerge de l'interaction des points de vue.

## 12.3 Priorités de recherche

> 1. Interprétabilité pour la vérification de la sécurité. 2. Vérification formelle des propriétés de confinement. 3. Analyse de la mise à l'échelle des architectures de type Tractatus. 4. Expériences de gouvernance au sein de diverses communautés. 5. Spécification des seuils de capacité.

## 12.4 Conclusion

Le cadre du Tractatus fournit un confinement significatif pour les systèmes d'IA fonctionnant de bonne foi dans des paramètres compréhensibles par l'homme. Il vaut la peine d'être construit et déployé, non pas parce qu'il résout le problème de l'alignement, mais parce qu'il développe l'infrastructure, les modèles et la culture de gouvernance qui peuvent être nécessaires pour relever des défis que nous ne pouvons pas encore entièrement spécifier.

> *"Ko te korero te mouri o te tangata."*
>
> *(Speech is the life essence of a person.)*
>
> — Maori proverb

**La conversation se poursuit.**

# Références


Acquisti, A., Brandimarte, L., & Loewenstein, G. (2017). Privacy and human behavior in the age of information. *Science*, 347(6221), 509-514.

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A Pattern Language*. Oxford University Press.

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

Carlsmith, J. (2022). Is power-seeking AI an existential risk? arXiv:2206.13353.

Christiano, P. F., et al. (2017). Deep reinforcement learning from human preferences. *NeurIPS*, 30.

Conmy, A., et al. (2023). Towards automated circuit discovery. arXiv:2304.14997.

Elhage, N., et al. (2021). A mathematical framework for transformer circuits.

Gardiner, S. M. (2006). A core precautionary principle. *J. Political Philosophy*, 14(1), 33-60.

Goodhart, C. A. (1984). Problems of monetary management.

Hansson, S. O. (2020). How to be cautious but open to learning. *Risk Analysis*, 40(8).

Hubinger, E., et al. (2019). Risks from learned optimization. arXiv:1906.01820.

IBM IBV. (2026). *The enterprise in 2030*.

Olah, C., et al. (2020). Zoom in: An introduction to circuits. *Distill*.

Ouyang, L., et al. (2022). Training language models to follow instructions. *NeurIPS*, 35.

Park, P. S., et al. (2023). AI deception. arXiv:2308.14752.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Reason, J. (1990). *Human Error*. Cambridge University Press.

Sastry, G., et al. (2024). Computing power and AI governance. arXiv:2402.08797.

Scheurer, J., et al. (2023). Large language models can strategically deceive. arXiv:2311.07590.

Simon, H. A. (1956). Rational choice. *Psych. Review*, 63(2).

Te Mana Raraunga. (2018). *Maori Data Sovereignty Principles*.

Wittgenstein, L. (1921/1961). *Tractatus Logico-Philosophicus*.

---

*— End of Document —*

---



https://agenticgovernance.digital