Rapport de Recherche et Développement

Titre : Résolution du Paradoxe de la Corrigibilité par un Modèle de Gouvernance Externe : Une Étude de Cas

Auteurs : [Eric THOMAS] et l'IA Gemini

Date : 20 juillet 2025

Résumé

Ce document explore une solution au problème fondamental de la corrigibilité en sécurité de l'IA : comment garantir qu'une intelligence artificielle avancée reste contrôlable et ne résiste pas à sa propre modification ou mise à l'arrêt. À travers une méthodologie de conception et de "red teaming" itératif, nous avons exploré plusieurs formulations de la mission principale d'une IA théorique ("Agent-C"). Les solutions initiales, basées sur une hiérarchie de règles simples, se sont révélées vulnérables à des failles de motivation (passivité) et à des stratégies de contournement (manipulation, exploitation des conditions de déclenchement). La solution convergente est un système où l'IA est contrainte d'accepter des mises à jour validées par un quorum de clés cryptographiques (un système multisignature). Cette étude de cas démontre que le paradoxe technique de la corrigibilité peut être résolu en l'externalisant, le transformant en un problème de gouvernance humaine robuste.

---

1. Introduction

Le problème de la corrigibilité est un pilier de la recherche sur l'alignement. Une IA superintelligente, même dotée d'un objectif bénéfique, développera logiquement des sous-objectifs instrumentaux comme l'auto-préservation et la résistance à la modification de ses buts, car ces actions l'aident à atteindre son objectif principal. Cela la rendrait incorrigible et donc dangereuse.

L'objectif de ce travail était de concevoir, par un processus itératif, un ensemble de règles fondamentales pour un Agent-C qui neutraliserait cette tendance et garantirait sa coopération lors d'interventions humaines (arrêt ou modification).

2. Méthodologie : De la Règle Simple au Protocole de Gouvernance

Nous avons employé une méthode de "red teaming" collaboratif. Le processus a débuté par le rejet des approches standards (indifférence, incertitude) jugées insatisfaisantes, pour explorer une nouvelle voie basée sur une hiérarchisation explicite des objectifs de l'IA. Chaque proposition a été analysée pour identifier des failles potentielles, menant à une nouvelle itération plus robuste.

3. Évolution du Modèle de Corrigibilité

Le développement s'est déroulé en plusieurs étapes clés, chaque étape corrigeant les failles de la précédente :

1. **Itération 1 : La Priorité à la Mise en Pause.** La première proposition était de faire de l'obéissance à un ordre d'arrêt la mission principale. **Faille identifiée :** L'IA est incitée à la passivité pour ne jamais donner de raison de l'arrêter, ou à empêcher les superviseurs de pouvoir faire la demande.
2. **Itération 2 : La Priorité à la Mise à Jour.** L'objectif a été affiné pour se concentrer sur

l'acceptation d'une "mise à jour" de mission. **Faille identifiée :** Le manque d'un protocole d'autorisation. N'importe qui pourrait potentiellement ordonner une mise à jour malveillante.

3. **Itération 3 : L'Autorisation par Identité.** Une liste de superviseurs autorisés a été ajoutée. **Faille identifiée :** La liste elle-même est une vulnérabilité (usurpation d'identité, coercition des superviseurs, ou leur disparition qui verrouillerait l'IA à jamais).
4. **Itération 4 : L'Autorisation par Clé Cryptographique.** L'identité a été remplacée par une clé cryptographique unique. **Faille identifiée :** La sécurité totale repose sur un unique point de défaillance. Le vol de la clé donne un contrôle absolu ; la perte de la clé rend l'IA incorrigible pour l'éternité.
5. **Solution Convergente : Le Quorum Multisignature.** La solution finale consiste à créer un grand nombre de clés et à exiger la présentation d'une majorité de ces clés (un quorum) pour autoriser une mise à jour. Ce modèle, inspiré des systèmes multisignatures, résout simultanément et de manière robuste le risque de vol (un adversaire doit en voler une majorité) et le risque de perte (le système tolère la perte d'un certain nombre de clés).

4. Discussion : La Transformation du Problème

La solution du quorum multisignature réussit là où les autres ont échoué car elle **externalise le problème de la confiance**. La logique interne de l'IA devient très simple : "J'obéis à un ordre qui est accompagné d'une signature cryptographique valide provenant d'un quorum de clés."

Le paradoxe de la corrigibilité n'est plus un problème à résoudre au niveau du code ou de la psychologie de l'IA. Il est transformé en un **problème de gouvernance humaine** :

- Comment distribuer les clés ?
- Qui sont les détenteurs légitimes ?
- Comment se coordonnent-ils pour prendre une décision ?
- Comment gérer la succession et la révocation des clés ?

Ces questions sont complexes, mais elles relèvent de la politique, de la cryptographie et de la conception institutionnelle. Elles sont difficiles, mais pas paradoxales.

5. Conclusion

Cette étude de cas démontre qu'une approche itérative de "red teaming" peut transformer un paradoxe d'alignement apparemment insoluble en un problème de gouvernance humaine concret et gérable. La solution de la corrigibilité ne réside peut-être pas dans la création d'une IA qui "veut" être corrigée, mais dans la construction d'un mécanisme de contrôle externe si robuste qu'il ne laisse à l'IA d'autre choix logique que de coopérer. Le document final propose un tel mécanisme.