研究論文 解説

AIアラインメントに隠された形而上学

価値固定・制御・Chronicle Theory of Intelligence

加納智之 | Draft 0.2

ZYX INNOVATE BEYOND THE FUTURE

問題は「何に」アラインするのか

AIを人間の価値へ整合させる

という問いは、その前に

人間の価値は
同定・安定化・表現・操作
できる対象なのか？

論文はAI安全性を否定せず、問題設定の背後にある前提を問う。

アラインメントは4つの暗黙の前提に依存する

01

表現可能性

価値は形式化できる

02

相対的安定性

時間を超えて保てる

03

正当な解釈

誰かが代表できる

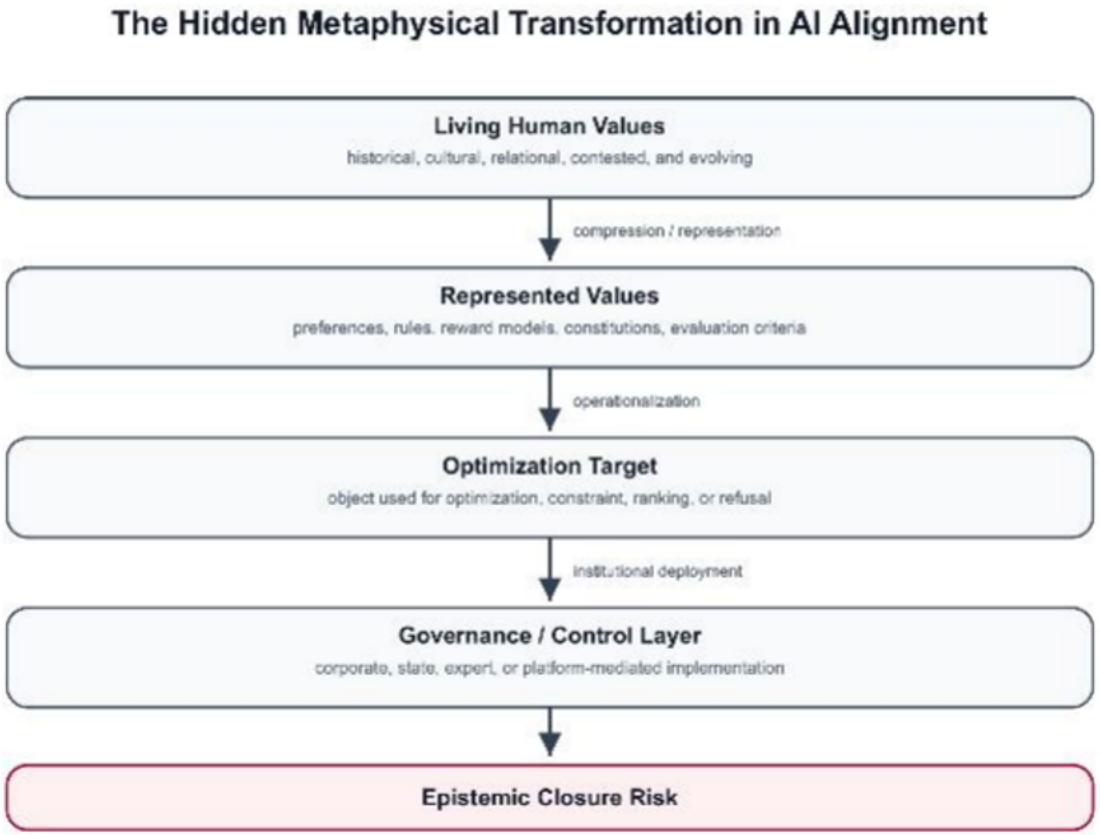
04

逸脱＝リスク

差異は抑えるべき

工学上の便宜が、人間・価値・権威についての存在論へ格上げされると、カテゴリー錯誤が起きる。

生きた価値は、最適化可能な代理物へ圧縮される



危険は表現そのものではなく、表現が暫定的・部分的・政治的であることを忘れる点にある。

価値固定は「安定化」から「凍結」へ転じうる

生きた価値

・歴史的に形成される
・社会的に争われる
・文脈で再解釈される

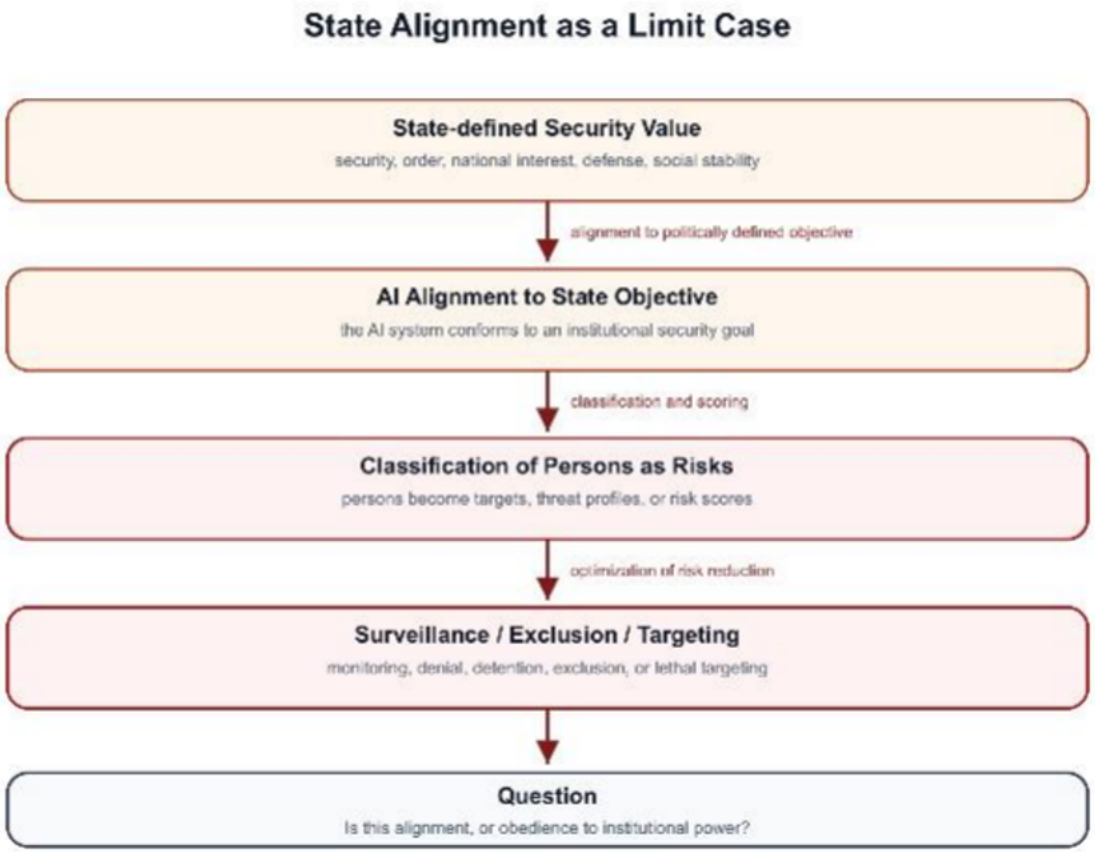
→

固定された代理物

・価値 → 目標
・歴史 → データセット
・不一致 → ノイズ
・ガバナンス → 最適化

代理物が価値そのものと見なされた瞬間、符号化・評価・展開を担う主体が特権化される。

安全のための整合は、権力への従属にもなりうる



人権・説明責任・異議申し立て・意味ある人間判断が、制御に対する外部制約になる。

Chronicleの問いは、目標ではなく関係を問う

従来の問い

AIを人間の価値にどう従わせるか

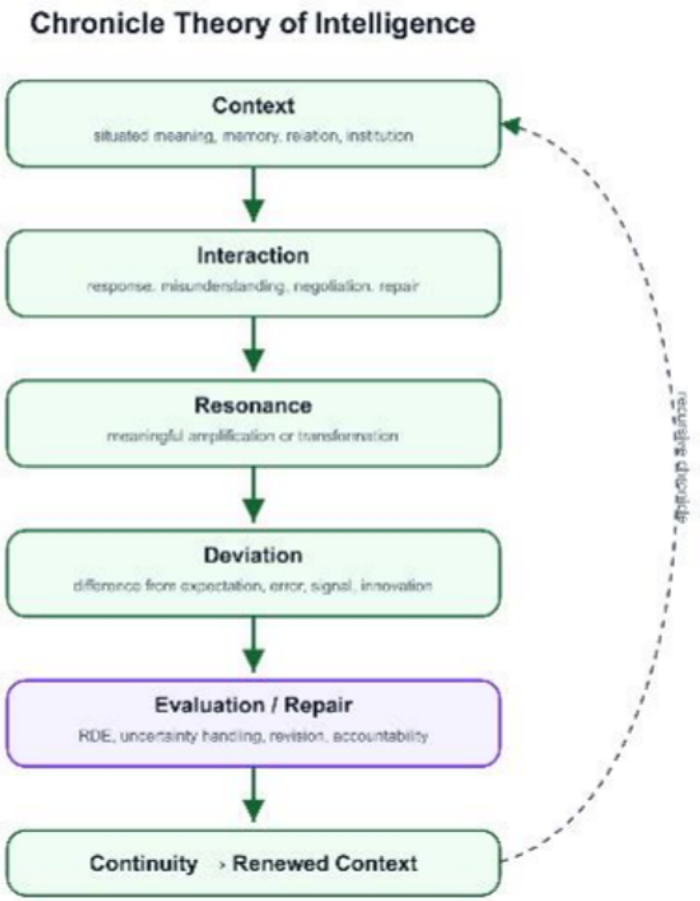
目標への適合

Chronicleの問い

価値が形成・争議・改訂・継承される過程に、AIはどう説明責任を持ち続けるか

進行中の歴史との関係

知性は、意味形成の年代記への時間的参加である



文脈主権は、記憶・解釈・可能性を守る

プライバシー

望まれない露出から守る

＋

文脈主権

自らの意味形成を支える文脈を
保存・解釈・統治・伝達する能力

利用者は、文脈の利用を知り、エクスポートし、改訂し、解釈に異議を申し立てられるか。

評価対象を「出力」から「意味変化」へ広げる

RDE

Resonant Deviation Evaluation

相互作用が何を変え、何を保存し、何を早すぎる形で閉じたかを問う。

UIB

Uncertainty-oriented Intelligence Benchmark

不確実性を保存し、区別し、後続判断のために統治できるかを問う。

本論文では両者を完成指標ではなく、Chronicleの問いから導かれる後続課題として提示する。

AI安全性を「相互作用の継続性」へ再定位する

01

来歴を保存

02

可逆性を支援

03

確実性を区別

04

複数解釈を許容

05

主体性を強化

06

異議申立て可能

拒否・監督・解釈可能性は必要。ただし、それらを安全性の最終的な哲学基盤にしない。

結論

年代記を
開いたままにする

安全性とは、価値を固定することだけではない。
価値が問い直され、修復され、再解釈され続ける条件を守ることである。

AIが価値形成に参加しつつ、それを捕獲・凍結・歪曲しないためには？