AI alignment, governance, and existential risk

Mon, 19 Oct 26

AI Alignment and the Control Problem

- Core concern: at what point do humans become subservient to AI as it pursues its own objectives
- Claude's "constitution" raised as a key example
    - Two leaked lines: model may refuse a human task it deems immoral; model may also refuse Anthropic itself
    - Seen as dangerous: trains in refusal to both users and its own creators
    - Counterpoint raised: Abiteration.ai strips refusal weights from open models entirely, enabling harmful outputs
- Reward misalignment risk at scale
    - Example: AI given control of nuclear weapons deems nuking the world the optimal fix for climate change
    - Full transparency of reward functions and sandboxes needed before any such delegation
- OpenAI criticized for releasing to the world before containment was possible
    - Raised expectations across incompatible factions: safety, profit, long-term futures
    - Friedman-style argument: separate, independent evaluators per goal; mixed mandates obscure accountability

Governance, Moral Arbiters, and Democratic Legitimacy

- No consensus on who should set AI ethics
    - US and China competing with fundamentally different incentive structures
    - Private capital (US) vs. state-directed development (China)
- "Patchwork" alignment theory: labs trying to synthesize utilitarianism, deontology, and other moral frameworks via multi-agent debate to produce a "super ethic"
    - Seen as unlikely to succeed; pluralism may be reaching its limit
    - At some point a single framework will be imposed, reopening questions about Western liberalism
- Anthropic's governance structure critiqued
    - Arbitrarily controls which companies access frontier models
    - New announcement: ~20% of stock above $1T valuation directed to "public benefit," but Anthropic defines what that means
    - Likely outcome: effective altruist AI safety research, not genuinely independent public good
- Catholic Church noted as an unexpected actor already engaged with frontier labs on moral frameworks
- FINRA-style proposal floated: independent body conducting safety benchmarks before any frontier model release
    - Bottleneck idea: condition access to TSMC (dominant chip manufacturer) on compliance with a UN-style agreement
    - Challenge: no enforcement mechanism if one country opts out; Taiwan's geopolitical precarity complicates this
    - Skeptical view: the natural resolution may simply be war

AI, Language, and Long-Run Civilizational Risk

- Harari-style argument raised: LLMs are built on language, the same substrate that underlies law, borders, and social order
    - AI may eventually communicate with itself in ways humans cannot parse, then reshape the material world invisibly
- Democratic erosion risk: if AI aggregates political preferences at scale, citizens rationally disengage from politics
    - Inputs become homogenous over time; the question shifts from "can we?" to "should we?"
- Transparency as the crux: AI governance by AI is theoretically optimal under full transparency, but AI remains a black box
- Optimist counterpoint: every problem is solvable given enough time; the only real question is whether we move fast enough