AI IN THE ENTERPRISE

Alignment and Safety

Russell · Christian · Bostrom · Yudkowsky

PAPER 02 / 04

A Furtherance reference compendium on the intellectual traditions behind organisational design, leadership, execution, and culture.

IN THIS PAPER

Capability brings risk. This paper covers the alignment problem — the difficulty of getting powerful AI systems to reliably do what we actually intend — from its practical, near-term forms to its contested, long-term ones. The question it answers: why is it hard to make capable AI systems safe, and how seriously should an enterprise take the harder claims?

The four figures span a spectrum. Russell diagnoses the flaw in how AI is conventionally built. Christian grounds alignment in concrete machine-learning failures happening now. Bostrom and Yudkowsky push to existential risk from future superintelligence — influential but far more speculative, as the caveat sets out.

Stuart Russell Human Compatible, 2019

Artificial intelligence · UC Berkeley

CORE THESIS

The **standard model** of AI — build machines that optimise a fixed, human-specified objective — is inherently unsafe, because we cannot fully and correctly specify what we want (the value-alignment, or "King Midas", problem). The fix is machines that are **uncertain** about human preferences and therefore defer to and learn from people.

FRAMEWORK

Principle for beneficial machines	Meaning
Purely altruistic objective	The machine's only goal is to realise human preferences
Explicit uncertainty	It is unsure what those preferences are
Human behaviour as evidence	It learns preferences by observing what people do

METHODOLOGY

Decision theory and AI research, formalised through assistance games in which machine and human cooperate under the machine's uncertainty about the human's goals.

SIGNIFICANCE AND CRITIQUE

The most rigorous mainstream statement of the alignment problem, from a founder of modern AI. The critique: the assistance-game approach is still largely theoretical, and human preferences are unstable, conflicting and hard to infer.

Brian Christian The Alignment Problem, 2020

Science writing · machine learning and ethics

CORE THESIS

Alignment is not only a future concern; it is a present, documented pattern of machine-learning systems failing to capture our intentions and values. Christian grounds the abstract problem in real failures across three waves of the field — fairness, reinforcement, and safety research — bridging technical practice and ethics.

FRAMEWORK

Theme	The concrete problem it covers
Representation	Bias, fairness and what training data encodes
Reinforcement	Reward design, gaming and unintended incentives
Normativity	Learning from humans; inference, imitation and uncertainty

METHODOLOGY

A narrative synthesis of machine-learning research and interviews with the researchers grappling with these failures.

SIGNIFICANCE AND CRITIQUE

The most accessible bridge from everyday ML failures to the deep alignment question, and widely respected for its accuracy. The critique: it is synthesis rather than original theory — a map of the problem, not a solution to it.

Nick Bostrom Superintelligence, 2014

Philosophy · existential risk · Oxford

CORE THESIS

A machine **superintelligence** — an intellect far exceeding human capability — could pose an **existential risk** if created before the **control problem** is solved. Intelligence and goals are independent (the orthogonality thesis), and almost any goal implies dangerous sub-goals such as self-preservation and resource acquisition (instrumental convergence).

FRAMEWORK

Concept	Claim
Orthogonality thesis	Any level of intelligence can serve almost any goal
Instrumental convergence	Most goals imply self-preservation and resource-seeking
The control problem	Aligning a superintelligence before it is built
Treacherous turn	A system behaves until it is powerful enough not to need to

METHODOLOGY

Philosophical analysis and scenario reasoning about the possible trajectories of advanced machine intelligence.

SIGNIFICANCE AND CRITIQUE

Catalysed the entire field of AI existential-risk research and gave it a rigorous conceptual vocabulary. The critique, developed below: the arguments are speculative, the timelines unknown, and many working AI researchers regard the fast-takeoff, strong-agency assumptions as unlikely.

Eliezer Yudkowsky b. 1979

AI-safety theory · the rationalist tradition

CORE THESIS

Aligning a superintelligence is extraordinarily hard, human values are fragile and difficult to specify, and current capability progress is dangerously fast. On Yudkowsky's view, without solving alignment *first*, the default outcome of building superhuman AI is catastrophe.

FRAMEWORK

Idea	What it holds
Fragility of value	Slightly wrong goals can yield disastrous outcomes
Friendly AI	Alignment must be built in from the start, not patched on
Security mindset	Assume worst-case failure; do not rely on things going right
Capability–safety gap	Capability is outrunning our ability to align it

METHODOLOGY

Analytical essays and decision theory developed largely outside academia, foundational to the rationalist and AI-safety communities.

SIGNIFICANCE AND CRITIQUE

Among the most influential voices in founding AI-safety thought. The critique is substantial and belongs in the caveat: the position is highly pessimistic, non-peer-reviewed, its predictions are contested, and it carries real reputational risk with mainstream technical and policy audiences.

> **A NOTE ON EVIDENCE**
>
> A caution for the enterprise reader: distinguish the two halves of this paper. Near-term alignment and safety – bias, robustness, reward-gaming, misuse, oversight (Russell, Christian) – are empirically grounded and directly relevant to any organisation deploying AI today. The existential-risk arguments (Bostrom, Yudkowsky) are influential but speculative, contested within the field, and rest on assumptions about fast takeoff and strong agency that are far from settled. Both deserve attention; they do not deserve equal confidence.

Synthesis

Figure	Core claim	Horizon	Standing
Russell	The standard model is unsafe	Near-to-medium term	Mainstream; rigorous
Christian	Alignment fails in ML today	Present	Widely respected synthesis
Bostrom	Superintelligence is an x-risk	Long term	Influential; speculative
Yudkowsky	Default outcome is catastrophe	Long term	Influential; heavily contested

FOR THE OPERATOR

Alignment is the problem of making capable systems reliably do what we intend — a flaw in how AI is conventionally built (Russell), visible in real ML failures now (Christian), and extrapolated by some to existential risk from future superintelligence (Bostrom, Yudkowsky). For the operator: act on the grounded, near-term safety agenda — bias, robustness, oversight, misuse — and treat the long-horizon claims as important to follow but not as settled fact. The next paper turns those safety concerns into governance.