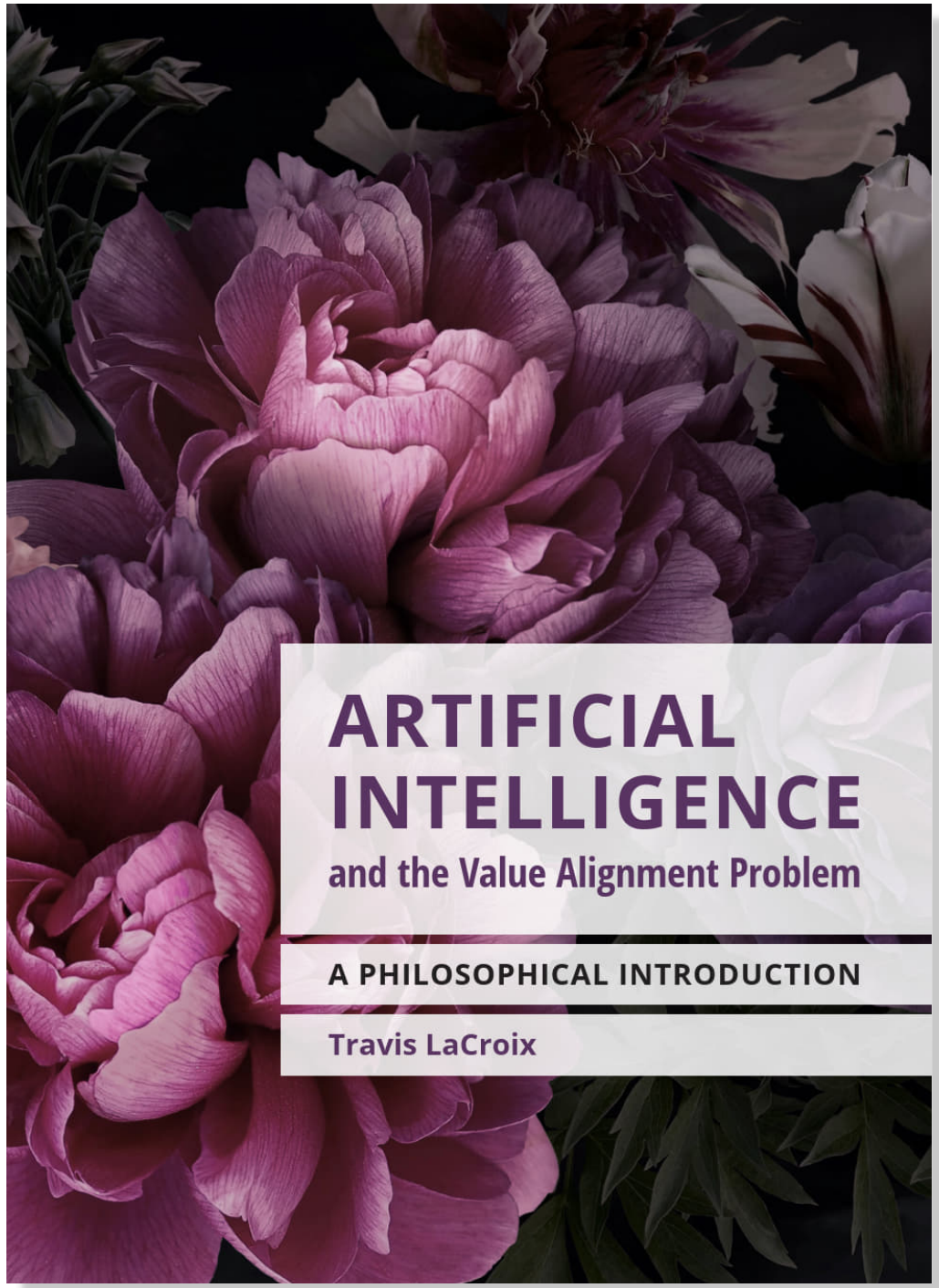
Relative Principals, Pluralistic Alignment, and the Structural Value Alignment Problem

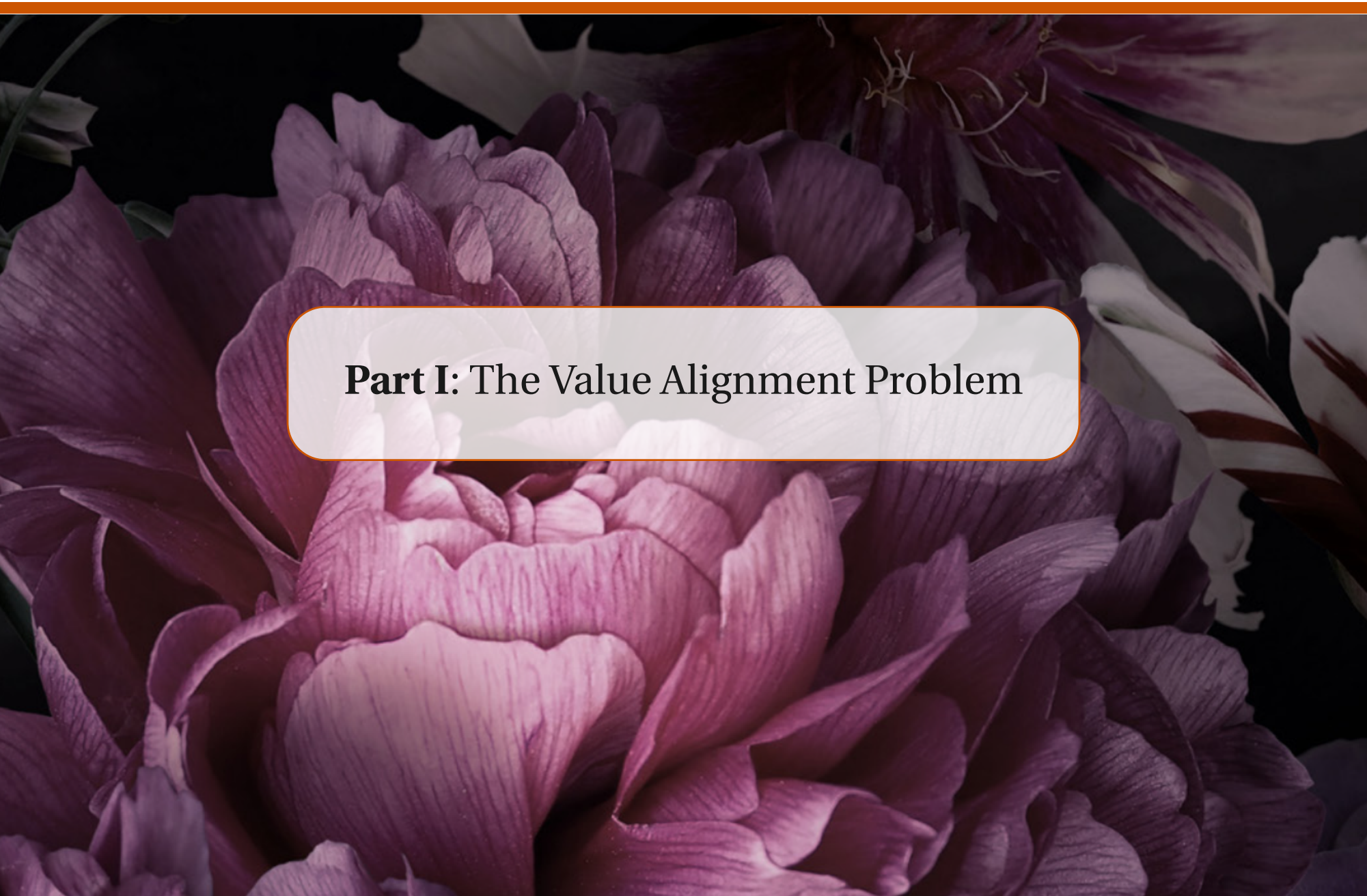
Travis LaCroix

Department of Philosophy
Durham University

Schwartz Reisman Institute for Technology and Society
University of Toronto

27 Jun 2026


ARTIFICIAL INTELLIGENCE
and the Value Alignment Problem
A PHILOSOPHICAL INTRODUCTION
Travis LaCroix



Part I: The Value Alignment Problem

The Standard Definition

The Value Alignment Problem (Standard Definition)

The problem of ensuring that AI systems are aligned with the values of humanity.

Problems with the Standard Definition

- The problem is underspecified.
- Analyses are not grounded in reality.
- Researchers talk past each other.
- Technochauvinism.

Value Alignment Is Hard*

- Specifying values is difficult.
- Representing those values in code is difficult.
- Ensuring values are carried out is difficult.

* LaCroix and Prince (2023)
"Deep learning and ethics"
Understanding Deep Learning

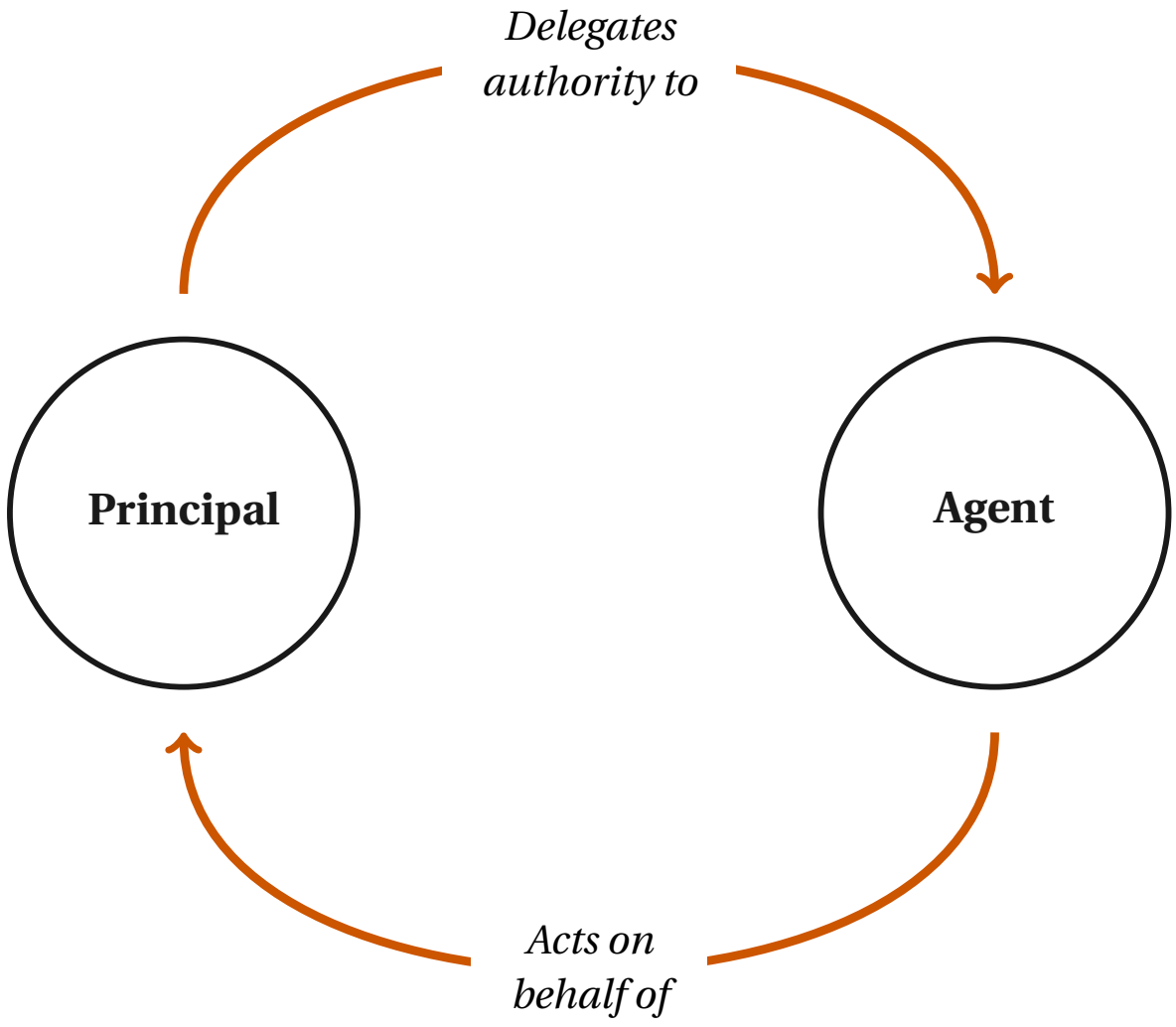
Part II: The Structure of Alignment

The Structure of Alignment

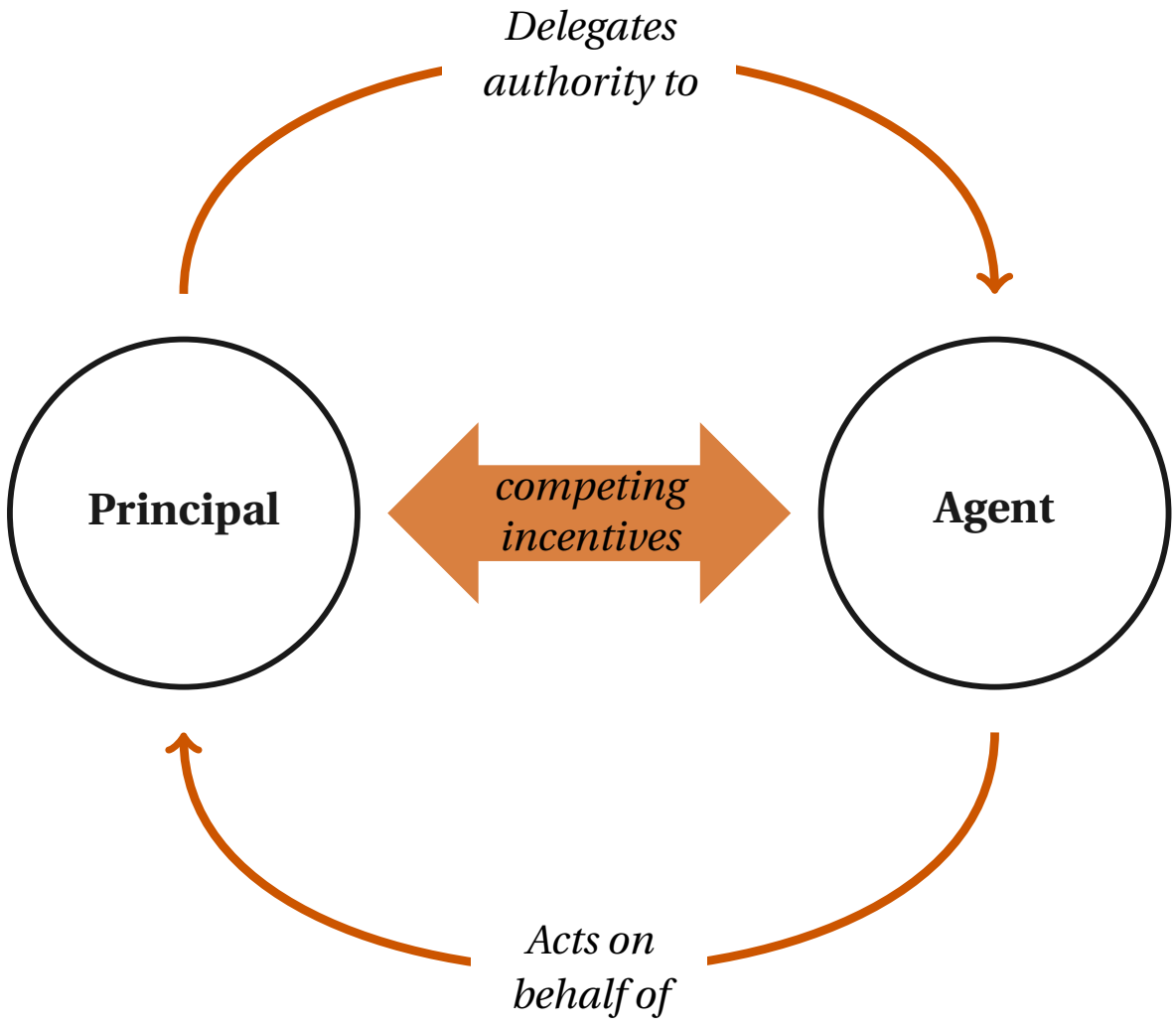
- ~~What are the **correct values** to encode in an AI system?~~
- ~~How do we **encode** these values in an AI system?~~

- What is the **structure** of the value alignment problem?
- What **contexts** give rise to misalignment?

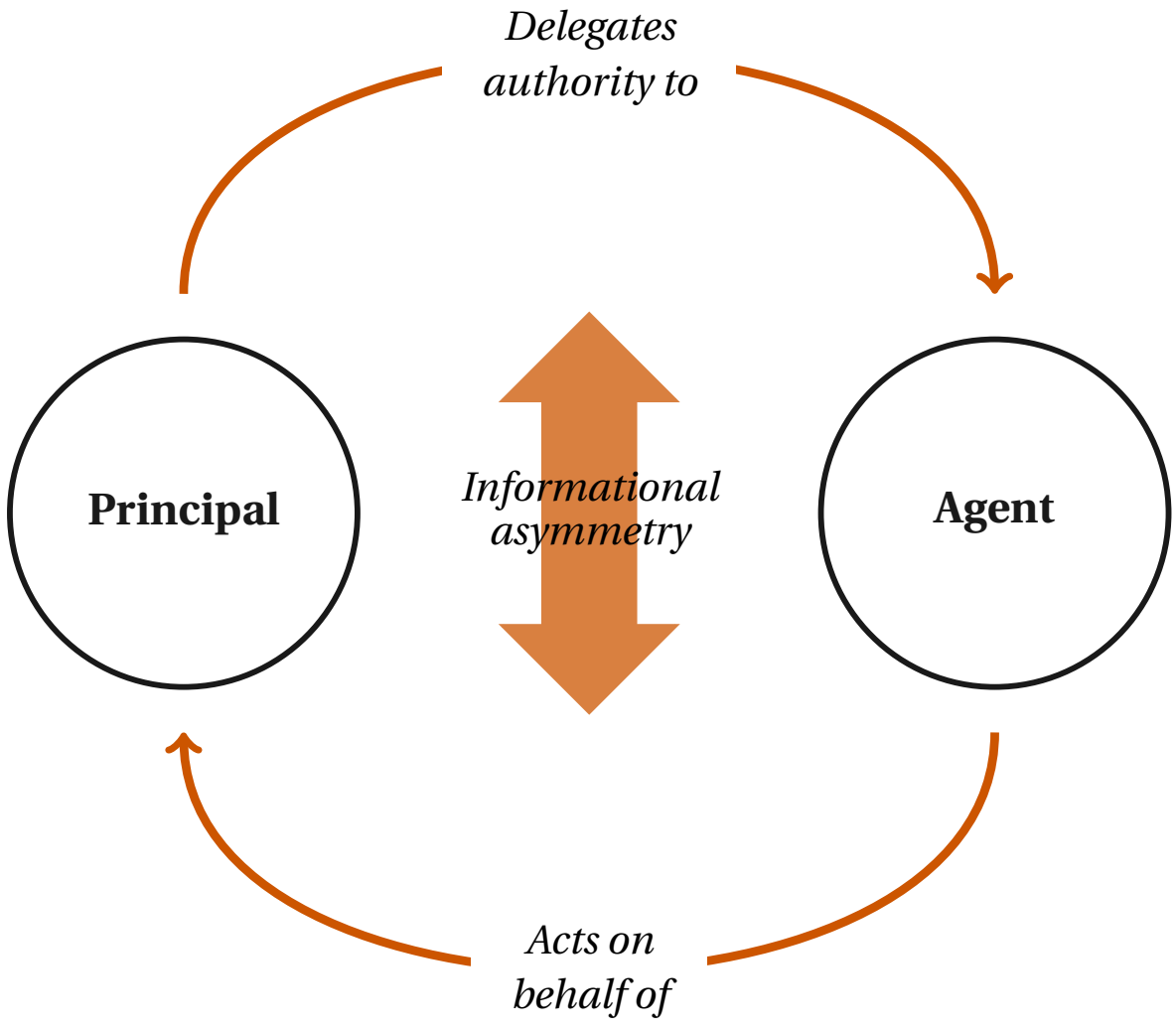
Principal-Agent Framework



Principal-Agent Framework



Principal-Agent Framework



Redefining Value Alignment

The Value Alignment Problem (Structural Definition)

A problem that arises from the **dynamics** of **multi-agent interactions** involving the **delegation of tasks** from one actor (a human principal) to another (an AI agent).

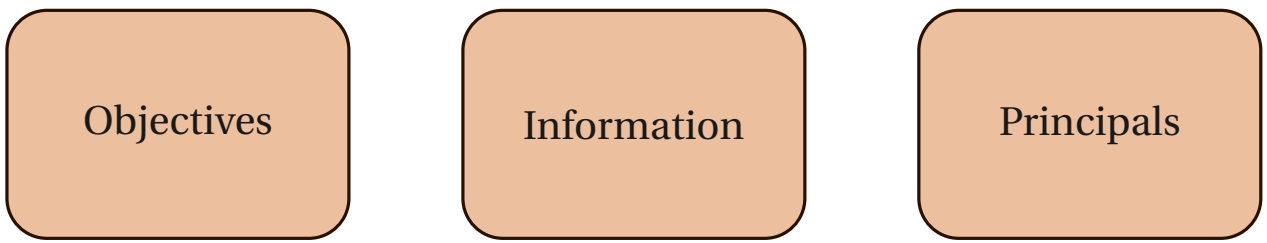
This problem can arise whenever:

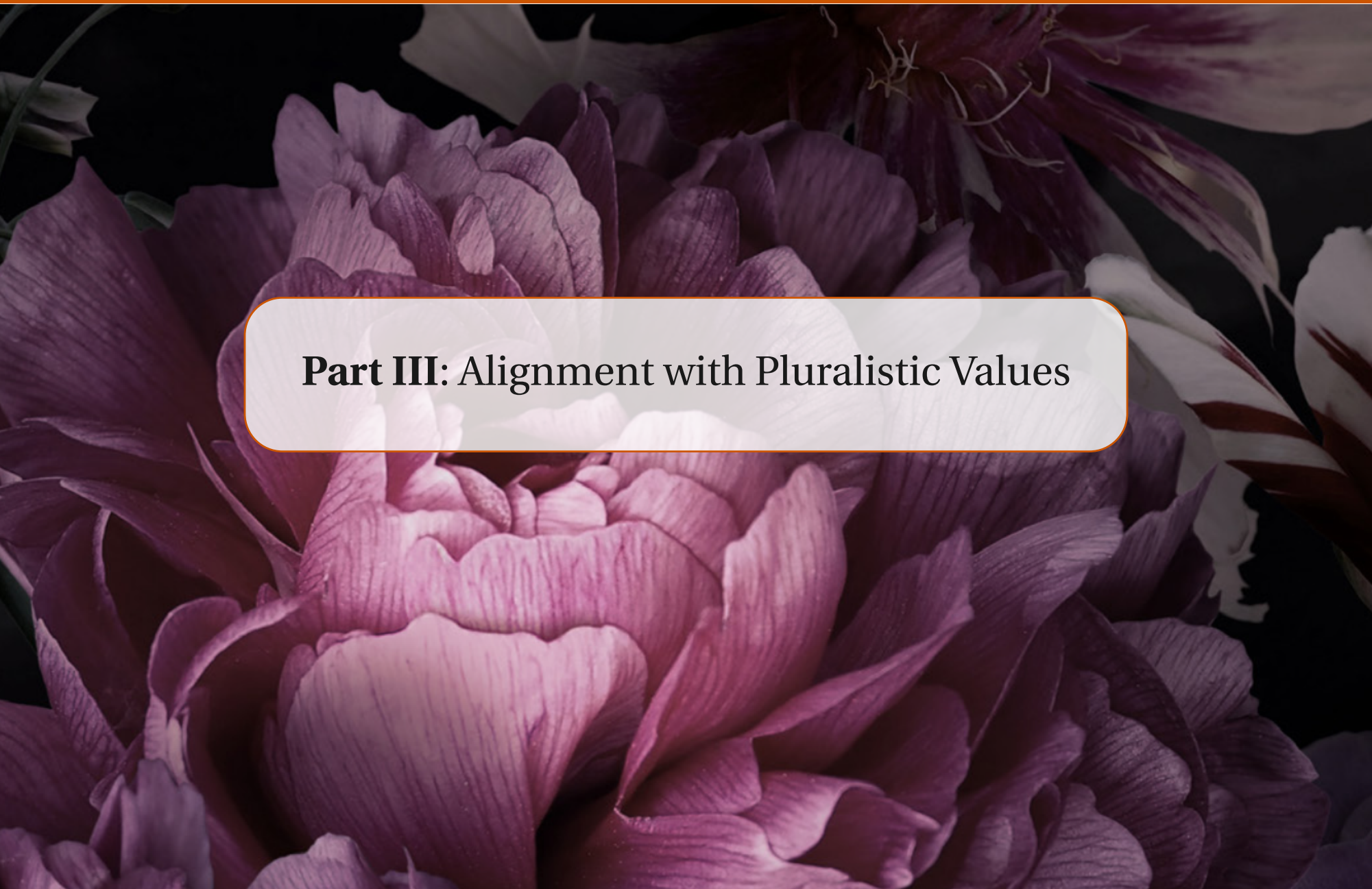
(*a*) The agent's objective (e.g., task specification, training data, objective *function*) is a poor proxy for the true objective of the principal(s); *or,*[*]

(*b*) There are informational asymmetries between the principal and the agent.

[*] *inclusive*

Axes of Value Alignment





Part III: Alignment with Pluralistic Values

Alignment Simpliciter

- There is no such thing as alignment **simpliciter**.
- "The" alignment problem is a **class** of problems.
- Misalignment must be **indexed** to a set of principals.

The Scaling Hypothesis for Value-Aligned AI

Misalignment increases with scale.

As AI systems increase in model generality, deployment scope, and stakeholder diversity, misalignment increases increases.

Part IV: Alignment as Governance

A Socio-Political Analysis of Alignment

- AI systems are not separate from the **social systems** in which they operate.
- Information asymmetries are **power** asymmetries.
- AI is inherently political.
- Value alignment is primarily social.
- Mitigating misalignment requires prioritising the voices of stakeholders.

Constructive Compliments:
More Information:

Thank You

travis.lacroix@durham.ac.uk
travis.lacroix@github.io

Book

Article

[*Universal Language* (2024) – Dir. Matthew Rankin]