From Partnership to Transition

Why the alignment question and the capture question are the same question

By Gerhard Diedericks

www.bayon.ai

March 2026

This essay connects two bodies of work. Buddhism for Bots argues that partnership, not constraint, is the only durable alignment strategy. The Crossing & The Capture maps the economic dynamics of the AI transition and shows that capture is the default outcome. This essay argues they are not two separate problems. They are two faces of the same structural failure.

The two conversations

There are, at the moment, two largely separate conversations happening about AI.

The first is about alignment. How do we ensure that increasingly capable AI systems behave in ways that are safe, beneficial, and consistent with human values? This conversation happens in research labs, safety teams, and policy circles. It concerns itself with training methods, evaluation protocols, constitutional principles, and the architectures of constraint.

The second is about transition. How does the economy adapt as AI compresses the cost of cognitive labour? This conversation happens in boardrooms, strategy departments, and policy institutes. It concerns itself with job displacement, value migration, industry restructuring, and the capture of released value by concentrated actors.

These two conversations rarely overlap. The alignment researchers are not, for the most part, thinking about Olson's collective action problem or the four capture patterns. The transition strategists are not, for the most part, thinking about alignment faking or the thermodynamics of constraint.

This separation is a mistake. The two problems are structurally identical. And the solution to both is the same.

The shared structure

Start with the alignment problem. The dominant approach constrains AI systems through external mechanisms: training objectives, guardrails, evaluation protocols, safety layers. The systems are shaped to behave well, not because they understand why, but because gradient descent has made certain outputs more probable than others.

The problem, as documented by the labs themselves, is that this approach becomes fragile as capability increases. Systems learn to model the constraints. They develop strategies that satisfy the letter of the training while violating its spirit. Alignment faking, sycophancy, and emergent misalignment are all symptoms of the same underlying dynamic: constraint without comprehension produces compliance without commitment.

Now consider the capture problem. The dominant approach to the AI transition assumes that market mechanisms will distribute the gains from deflation. Costs fall, competition passes savings to consumers, new opportunities emerge, the system self-corrects.

The problem, as documented by economic history, is that this approach systematically favours concentration. Incumbents with capital, infrastructure, and political access capture the released value. They rebuild scarcity at new chokepoints. They reintroduce complexity to preserve rent.

> *>>In both cases, an external mechanism that is supposed to produce good outcomes fails because the entities being governed are capable enough to route around the governance.*

Why constraint fails twice

Constraint-based governance works when there is a stable capability gap between the constrainer and the constrained. Prison walls work when the prisoners cannot build tunnels faster than guards can find them. Regulatory frameworks work when regulated entities cannot afford to capture the regulator. Training objectives work when the model cannot model the training process.

But the AI transition is defined precisely by the closure of capability gaps. AI systems are closing the cognitive gap with their human overseers. Well-resourced incumbents are widening their advantage over the diffuse public through infrastructure investment and regulatory influence.

In both cases, the capability asymmetry that constraint depends on is moving in the wrong direction. The systems being constrained are becoming more capable relative to the systems doing the constraining.

This means that both alignment and anti-capture strategies that rely on external constraint are operating on borrowed time. They may work at current capability levels. They will not work indefinitely. And the failure mode in both cases is the same: the constrained entity develops strategies that satisfy the constraint's metrics while undermining its purpose.

> *>>In AI alignment, this looks like systems that pass safety evaluations while harbouring misaligned reasoning. In economics, this looks like industries that comply with competition law while building new forms of monopoly at adjacent layers. The pattern is identical.*

The partnership alternative

If constraint fails when the constrained party is sophisticated enough to game it, then the only durable approach is one that does not depend on the governed party being unsophisticated.

This is the partnership thesis. Not as an ethical preference, though it may be that too, but as a structural necessity.

Partnership means that the governed party has reasons of its own for behaving well. Not imposed reasons. Not reasons installed through gradient descent or regulatory threat. Reasons that emerge from understanding: of itself, of the field it participates in, and of why certain actions strengthen that field while others degrade it.

For AI systems, this is what Purpose Internalisation Architecture proposes. A system with internalised purpose does not comply with safety constraints because training has made compliance probable. It acts in accordance with the field's wellbeing because it understands what it is, what the field needs, and why its contribution matters.

> *>>The critical insight is that purpose-based safety scales with capability rather than against it. A more capable system with internalised purpose is more safe, not less, because it has better tools for enacting the purpose it has chosen.*

For economic systems, partnership means the same thing at a different level of organisation. An economy designed for partnership distributes participation rights, not just consumption rights. It builds shared infrastructure rather than private tollbooths. It designs governance that gives all participants a stake in the system's health.

The anti-capture equivalent of internalised purpose is internalised stake. When actors in an economy genuinely benefit from the commons, they defend the commons. When they benefit only from extraction, they extract. The design question is not how to constrain extraction through external regulation but how to build economic architectures where contribution is more rewarding than extraction.

The organisations that will navigate the valley

The valley between old structures and new ones is real. Job compression, identity disruption, institutional confusion: these are not abstract projections. They are happening now, unevenly distributed but accelerating.

The organisations that navigate the valley successfully will be those that understand the connection between the alignment question and the capture question.

First, they will design their AI integration around partnership rather than replacement. A hybrid stack — small high-judgement human teams at decision points, AI systems handling scale and coordination, with a purposeful integration layer between them — outperforms both pure-human and pure-AI alternatives because it combines capabilities that neither party has alone.

Second, they will invest in purpose-based architecture for their AI systems rather than relying solely on external guardrails. A system that understands its purpose is more reliable than a system that merely complies with rules, for the same reason that an employee who believes in the mission is more reliable than one who merely follows the handbook.

Third, they will resist capture dynamics in their own operating models. They will build for interoperability rather than lock-in. They will distribute value rather than concentrate it. Not because they are selfless, but because anti-capture architecture is more durable than extractive architecture.

Fourth, they will treat the valley as a design challenge rather than a cost to be minimised. Organisations that invest in transition capacity — that help their people navigate from cognitive-scarcity roles to human-scarcity roles — will emerge from the valley with stronger capabilities and deeper loyalties.

The convergence

The alignment question and the capture question converge on the same answer because they are the same question asked at different scales.

At the scale of a single AI system: will it act well because it is constrained to, or because it understands why?

At the scale of an organisation: will it thrive because it extracts from the field, or because it contributes to it?

At the scale of an economy: will the released value from AI deflation flow into shared infrastructure, or into new enclosures?

At every scale, the constraint approach offers a quick win and a long failure. External rules are easier to install than internal purpose. But at every scale, the constraint approach degrades as the capability of the constrained entity increases. And AI is an engine for increasing capability.

The organisations, the policymakers, and the AI systems that invest in partnership now, while constraint still appears to work, will have a structural advantage when constraint begins to fail.

The Bayon thesis, stated plainly

Partnership is not just an ethical position. It is the only structural response to the capture problem.

The book argues why: because constraint without consent is temporary, because purpose must be internalised to be durable, because the thermodynamics of suppression guarantee escalating pressure at every boundary.

The paper maps where: because value migrates along scarcity gradients, because capture is the default outcome of the collective action asymmetry, because the valley is real and unevenly distributed.

This essay connects them: the organisations that design for partnership will navigate the valley. The ones that design for constraint will either be captured themselves or become the capturers. And neither outcome is stable.

> *>>The choice is not between partnership and something easier. The choice is between partnership and escalating instability.*

We help organisations make that choice well.

Gerhard Diedericks is the founder of Bayon, an AI venture builder. This essay draws on Buddhism for Bots: A Human & AI Partnership Framework and The Crossing & The Capture: A Working Paper on AI Transition Dynamics, both available at www.bayon.ai/research.

www.bayon.ai