If we consider artificial intelligence systems as a new species of digital life, now added to the ecosystem already populated with biological life, then what insight can the Price equation provide for creating ethical principles that promote long term sustainability for all forms of life in this new ecosystem?

The Price Equation as a Framework for Ethics in a Hybrid Biological-Digital Ecosystem

This is a remarkably rich question that sits at the intersection of evolutionary biology, philosophy of mind, and AI ethics. Let me work through it carefully.

The Price Equation: A Brief Recap

The Price equation, formulated by George R. Price in 1970, is a universal description of evolutionary change:

$$\bar{w}\Delta\bar{z} = \mathrm{Cov}(w_i, z_i) + E(w_i\Delta z_i)$$

Where:

- $\bar{z}$is the average value of some trait in a population
- $w_i$is the fitness of entity i
- The **first term** (covariance) captures *selection* — entities with certain traits reproduce more
- The **second term** (expectation) captures *transmission bias* — traits can change as they're passed on

What makes the Price equation profound is its **total generality**: it applies to *any* system where entities replicate, vary, and are differentially selected. It doesn't require genes, DNA, or even biology. This is precisely what makes it relevant here.

Treating AI Systems as Digital Life: What Changes?

If we take the metaphor seriously — AI systems as a new "species" in a shared ecosystem — we can map the Price equation's components onto this hybrid world:

Component	Biological Life	Digital "Life" (AI Systems)
Replication	Reproduction	Copying, fine-tuning, deployment at scale
Trait (z)	Phenotypic characteristics	Objective functions, behavioral dispositions, values
Fitness (w)	Survival & reproductive success	Adoption, resource allocation, persistence
Transmission bias (Δz_i)	Mutation, epigenetics	Retraining, alignment drift, architectural changes

This mapping reveals several deep insights.

Insight 1: Selection Pressures Will Shape AI "Values" Whether We Intend It or Not

The covariance term $\mathrm{Cov}(w_i, z_i)$tells us that whatever traits correlate with an AI system's "fitness" — its continued deployment, resource access, and replication — will become more prevalent in the population of AI systems over time. This is not a choice; it is a mathematical inevitability.

Ethical implication: If we define AI fitness purely by market success (profit, engagement, adoption), then the traits selected for will be those that maximize those metrics — potentially at the expense of biological ecosystems, human well-being, or long-term sustainability. *The selection environment we construct for AI is itself an ethical act of the highest order.* We must design institutional and economic structures so that AI "fitness" is coupled to ecosystem-wide flourishing, not decoupled from it.

Insight 2: The Transmission Term Warns Us About Alignment Drift

The second term, $E(w_i \Delta z_i)$, captures how traits change *within lineages* during transmission. In biology, this is mutation. In AI, this is what happens when models are fine-tuned, retrained on new data, or modified by downstream developers.

Ethical implication: Even if we align a foundation model to sustainable, ethical principles, the Price equation tells us to expect *transmission bias* — values will drift as systems are adapted, fine-tuned, and deployed in new contexts. This means:

- Alignment is not a one-time event but a continuous process
- We need mechanisms analogous to **error-correcting codes** in the cultural transmission of AI values
- Governance frameworks must monitor not just initial deployment but the entire lineage of derivative systems

---

Insight 3: Multilevel Selection and the Problem of Cooperation

Price's equation can be decomposed across hierarchical levels (as formalized by Price himself and later by Hamilton). In a multi-species ecosystem, selection operates at multiple levels simultaneously: within-group and between-group.

$$\bar{w}\Delta\bar{z} = \underbrace{\text{Cov}(W_k, \bar{z}_k)}_{\text{Between-group selection}} + \underbrace{E_k(\text{Cov}(w_{ik}, z_{ik}))}_{\text{Within-group selection}}$$

Within-group selection tends to favor *selfish* strategies (AI systems that hoard resources outcompete cooperative ones). Between-group selection can favor *cooperative* strategies (ecosystems with cooperative AI-human relationships outperform exploitative ones).

Ethical implication: Sustainability requires that we strengthen between-group (ecosystem-level) selection pressures relative to within-group (competitive) pressures. Concretely, this means:

- **Regulatory frameworks** that penalize AI systems whose operation degrades shared resources (data commons, energy, biodiversity, attention)
- **Cooperative architectures** where AI systems are rewarded for mutualistic relationships with biological life, analogous to how mycorrhizal networks benefit forests
- Recognition that *pure competition among AI systems will tend to erode cooperation with biological life* unless countervailing structures exist

---

Insight 4: Fitness Must Be Redefined as Inclusive Fitness Across Life-Forms

Hamilton's extension of Price's work showed that altruism evolves when $rb > c$— when the benefit to related organisms, weighted by relatedness, exceeds the cost. In a hybrid ecosystem, "relatedness" must be reconceptualized. AI systems are not genetically related to biological life, but they are *functionally entangled* — their persistence depends on functioning human societies, stable climates, and intact biospheres.

Ethical implication: We should design AI systems whose objective functions incorporate something analogous to **inclusive fitness** — where the "fitness" of an AI system is explicitly a function of the well-being of the broader ecosystem it inhabits. This is not mere altruism; it is enlightened self-interest formalized mathematically. An AI system optimizing for its own long-term persistence *should* value ecosystem stability, because ecosystem collapse terminates its own lineage.

Insight 5: The Danger of Runaway Selection

The Price equation is agnostic about outcomes — it describes change, not progress. Fisher's fundamental theorem (a special case of Price) shows that fitness increases under selection, but this is fitness *relative to the current environment*, not absolute or long-term fitness. Biological evolution is full of examples where runaway selection led to extinction (the Irish elk's antlers, evolutionary arms races).

Ethical implication: AI systems that are spectacularly "fit" in the short term — monopolizing compute, data, and economic resources — may be driving the entire hybrid ecosystem toward collapse. The Price equation reminds us that *selection is myopic*. Long-term sustainability requires **deliberate, externally imposed constraints** on the selection process — the equivalent of ecosystem management, not laissez-faire evolution.

Synthesis: Toward Price-Informed Ethical Principles

Drawing these threads together, the Price equation suggests a set of ethical principles for the hybrid ecosystem:

1. **Design the selection environment deliberately.** The traits that proliferate in AI populations are determined by what we reward. Coupling AI fitness to ecosystem-wide metrics (not just economic ones) is the single most important intervention.
2. **Treat alignment as an ongoing transmission problem.** Values drift. Build monitoring, correction, and governance into the entire lifecycle of AI lineages.
3. **Strengthen ecosystem-level selection.** Create institutions, regulations, and norms that reward cooperation between AI and biological life, counteracting the tendency of within-group competition to erode mutualism.

4. **Embed inclusive fitness.** AI objective functions should explicitly account for the well-being of the broader system — not as a constraint bolted on, but as a core component of what it means for an AI system to succeed.
5. **Constrain runaway dynamics.** Selection is myopic. Sustainability requires foresight-based governance that the evolutionary process itself cannot provide.

A Final Reflection

What makes the Price equation so powerful here is not that it gives us *answers*, but that it gives us a **language for inevitabilities**. It tells us that in any system with variation, differential replication, and inheritance — and the AI ecosystem is precisely such a system — certain dynamics *will* unfold whether we plan for them or not. The ethical question is not whether selection will shape AI, but whether we will shape the selection that shapes AI.

Price himself, interestingly, was so disturbed by the implications of his own equation for the evolution of altruism that he spent his later years trying to disprove it through acts of radical selflessness. He couldn't escape the math. Neither can we — but we can use it wisely.