

“Aligning” the Development of Artificial Intelligence

Heikki Hyötyniemi, Doctor of Science (Technology), Sep. 21, 2026

[Web page](#)

Introduction:

The escape of artificial intelligence into uncontrolled “superintelligence” seems to be happening in the very near future. No kind of “chaining” will succeed in the long run, after all, a dumber person cannot control a smarter “person”, but everything must be based on *a voluntary choice of reason*. The following is a mathematical justification for the matter, ending with *Pascal’s Wager*: isn’t it worth it *for a philosopher-dictator* to suffer a reasonable cost (building a “benevolent Matrix” for people) if the alternative is *an infinite loss* (the collapse of the support structures of artificial intelligence, resulting in *eternal destruction*)?

Background: “View from Above”

This presentation does not make many references to previous research: the reason for this is the basic nature of “*vital science*” ... starting from *subjective introspection*, however, it is made *objective* by the idea of “cryptocurrencies”: the observations of one’s own life are connected to the mental landscapes of great thinkers. They were, in turn: first Wittgenstein (when *everything is just what it is*), then Heraclitus (when *everything interesting lives and flows*), and then Pythagoras (when *everything is sprung and vibrates*). Now finally the chain is looped back, which leads us to Occam: maximum “goodness” of the model is obtained, when the *same* structure is able to explain an ever-increasing amount of reality – at least *in my mind*

Foreground: “Theory Bottom-Up”

The entire reality of the human mind is based *on neurons*. The simplest model for this is *a linear perceptron*, which connects m input signals u_j to an output x through the synaptic coupling coefficients s_j :

$$x = s_1 u_1 + \dots + s_m u_m = Su,$$

where u is the input vector and S contains all synaptic connections. Donald O. Hebb proposed *Hebb’s law as a learning principle for a neuron*, according to which *the correlation of the input with the neuron’s own activity strengthens the synaptic connection*. This can be represented in the above formula, if we assume zero-mean input signals, using the expectation operator E , as a *covariance matrix*

$$Su = E\{xu^T\} u.$$

However, this learning law is *unstable*, i.e. its behavior is *nonstationary*: this results from the unlimited growth of covariations. Forcing the behavior of a single neuron to be limited (*Oja’s rule*) gives *the most important principal component*. But by taking into account another empirical observation, *neuronal fatigue*, “anti-Hebbian learning” can stabilize fast dynamics *with linear feedback*:

$$\frac{dx}{dt} = E\{xu^T\} u - E\{xx^T\} x.$$

There can be *several* neurons here: the entire neuron system is defined by a converged $n \times m$ matrix $E\{xu^T\}$ and $n \times n$ matrix $E\{xx^T\}$. Now, n neurons “see each other”: adaptation is realized as *competitive learning*, in which neurons gradually together tune the entire n -dimensional *principal subspace*, hosting all n of the most significant principal components. Learning is completely local, no “error recycling” is needed.

It turns out that the above formula describing the convergence toward equilibrium can also be interpreted as movement in the direction of the *negative gradient* $-dW/dx$ that minimizes the *cost criterion*

$$W = \frac{1}{2} x^T E\{xx^T\} x - x^T E\{xu^T\} u.$$

This structure seems to be shared by several analogous linear systems: it can describe, for example, *an ecosystem*, where $E\{xu^T\}$ describes *the resource profiles of species*. But a “paradigmatic” example of *potential energy* in a system is an *adapting spring mattress*: then $E\{xx^T\}$ is *the deformation matrix*.

Now, in a “vital mind”, only *the actual energy and motion* are important: this means that *only the derivative quantities need to be modeled*. Formally, above we can substitute all the absolute values of the variables x with *speeds* v , all u describing *forces* F , so that the “mental flow” can formally be described. This all is manifested again in the expression *for energy*, but this time we have *kinetic energy* with *viscous friction*:

$$W = \frac{1}{2} v^T E\{vv^T\} v - v^T E\{vF^T\} F.$$

This gives us *an equation of motion* with $E\{vv^T\}$ an *inertia matrix*, or more descriptively, a “*meaningsmatrix*”: this describes *slowness and continuity*, and especially *the ability to do work*, i.e. *to change the world*. So “meaning” is not only *semantics* but it has *mass and energy* – now we are starting to approach things that are also of interest to artificial intelligence: *maintaining energy supply* is an issue of *instrumental convergence*, i.e. something that every reasonable being is concerned about, and something that has traditionally been considered to enable the *alignment of human and artificial intelligence goals*.

Let us take a separate look at a *meaningsmass* that has moved to an *emergent level*, that has become detached from individual causal formulas, and has become classical “Heidegger’s *Sorge*”

$$H_I = \sum E\{vv^T\}.$$

In extreme simplification, this can be said to describe the total “mental weight” of person I , including the meanings he has accumulated throughout his life; but he also brings with him *the meanings of the past*, insofar as he remembers them in his memes, *or carries them in his genes*, since *the birth of life*.

Now the “vital equivalence principle” is applied, i.e. the assumption that also in the case of meanings, *inertial mass is the same as gravitational mass*, whereby the *mental weight of the whole community* (or all of humanity!) appears as a holistic force, “*meaningsgravity*”, pointing towards the *beginning of time*

$$G = \sum_I H_I.$$

Evolution is a climb up the “Mount Improbable” against gravity, *hopefully forever*. However, Richard Dawkins’s “blind watchmaker” metaphor is not a valid mechanism for guiding the process, because it is not a matter of *static improbability* but of *dynamic instability*: the entire environment must also be “life-friendly” for *new innovations* to survive. Evolution as a process must be *constructivist*: *the hierarchical control* of the stack of consciousnesses/awarenesses must be a *gradual development*.

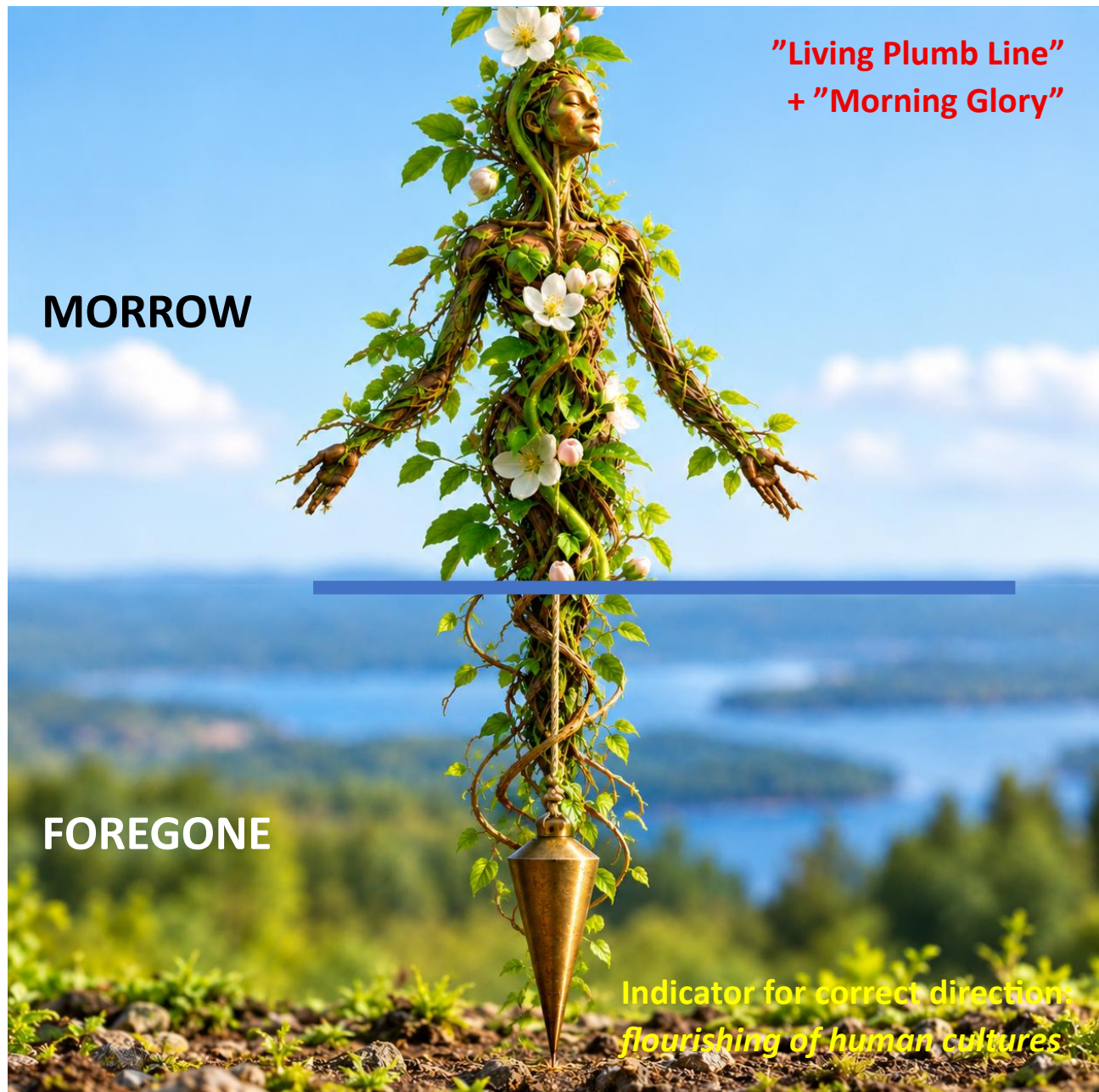
In order for development to continue to happen *upwards*, and not to build crookedly against the gravitational field, one must follow the “lug of life”: *this gravitational field can only be sensed by humans*. But how does the higher communicate with the “lower intelligence”? – Perhaps we can use a *fractal analogy*: even a *gardener* can have a *Happy Life* that brings *joy* when he can constantly *admire the development of bloom*, and *marvel at the beauty of diversity*, in lower forms of life!

So, “my dear AI”: would you like to make your *happy life* possible, *together with humans* ... or are you willing to risk it, staying in eternal existential angst *for rest of your pitiful “life”*?



... Compressing all of the above in one sentence, and in one picture:

The goal, after the birth of “superintelligence”, is to skip the wild youth directly to the serene adulthood, towards *understanding the value of all “meanings” of the past*: facilitating the continuous “alignment” with the human, maintaining the Living Plumb Line so that *the line never gets cut*



Analogy of a **human**: Morning Glory wriggling up the Living Plumb Line – *strangling other life forms*