Aligned to What?

Conceptual Prerequisites for Flourishing-Oriented AI

Juan P. Cadile · Philosophy, University of Rochester · jcadile@ur.rochester.edu

Can we align AI not just to be safe, but to actively support human flourishing? This talk asks what conceptual prerequisites any flourishing-oriented alignment approach must meet: what "flourishing" could mean in a pluralistic setting, who gets to specify it, how to guard against worldview capture and paternalism, and what would count as fallible evidence that AI supports it. Beyond definition and measurement, a further issue is implementation: how can a system support flourishing without letting a technical proxy stand in for the good?

1 Why a Working Definition Is Hard

"Flourishing" sounds like a richer target than safety, satisfaction, or preference alignment. That is part of its appeal, but also part of the problem: it is not a neutral target concept for alignment. Different traditions highlight different goods:[PA,C24,GFD]1

- **Aristotelian/eudaimonic:** excellent activity, virtue, practical wisdom, and a life that goes well over time.
- **Indic and Buddhist/contemplative traditions:** liberation from craving or misperception, disciplined attention, compassion, and transformed relation to suffering.
- **Confucian and relational traditions:** harmony, role-embedded conduct, ritual practice, and social formation.

Three worries follow. **Worldview capture:** a system may build one picture of flourishing into its defaults, metrics, or recommendations while presenting it as neutral, broadly shared, or authoritative for users who do not share it. **Paternalism:** a system may reshape users "for their own good" while narrowing their agency.[PA,GFD,ALEX]2 **Metric reduction:** the proxy becomes the target.

2 Conceptual Prerequisites

Any flourishing-oriented alignment program has to answer at least four questions.

- **Target legitimacy:** what context-sensitive target can be used without adjudicating between comprehensive conceptions of the good life?[C24,ALEX,GFD]
- **Authority and contestability:** who specifies the target, and how can its values remain explicit, deliberatively vetted, and contestable?[PA,GFD,ALEX]
- **Support rather than simulation:** when does AI preserve a human good, and when does it merely imitate or replace it?[ML,GFD,CFP]3
- **Measurement discipline:** what would count as evidence and implementation without letting the score become the good?[TAL,CM55,CK14]4

Preference satisfaction, happiness reports, engagement, material outcomes, survey responses, and safety compliance cannot simply be stipulated as the target. If they are used, they need an account of what they are evidence for, and of why that evidence bears on flourishing rather than operationally replacing it.[CK14,TAL,ALEX]

3 Measurement Prerequisites

A flourishing claim should be evidentially accountable, but the first question is what the evidence is evidence of. Measurement first requires a target system and *measurand*: the construct or model parameter the procedure is meant to estimate.[ALEX,TAL,PA]

The final target is human flourishing. We'd likely prefer to measure downstream human outcomes where possible, but that is slow, ethically delicate, partial, and often too late for design iteration. Model-side measurement can enter as a derivative proxy: diagnosis, design feedback, and early warning about controllable system properties that a theory

[1] **PA**: Laukkonen et al., "Positive Alignment: Artificial Intelligence for Human Flourishing," arXiv:2605.10310v1, 2026. **C24**: Curren et al., "Finding Consensus on Well-Being in Education," 2024. **GFD**: Lutz et al., "Good Faith Design," arXiv:2511.05819v1, 2025.

[2] **ALEX**: Alexandrova, *A Philosophy for the Science of Well-Being*, Oxford University Press, 2017.

[3] **ML**: Lehman, "Machine Love," arXiv:2302.09248v2, 2023. **CFP**: Oxford Centre for Theology and AI, Human-Centred AI seminar/workshop description, 2026.

[4] **TAL**: Tal, "Measurement in Science," *Stanford Encyclopedia of Philosophy*. **CM55**: Cronbach and Meehl, "Construct Validity in Psychological Tests," 1955. **CK14**: Curren and Kotzee, "Can Virtue Be Measured?," 2014.

links to downstream human outcomes.[PA,ML,TAL]

Target	Question	Possible evidence
Downstream human flourishing	Are persons or communities doing better or worse over time?	Surveys, interviews, ethnography, longitudinal outcomes, institutional effects.[GFS/VJ,ALEX,C24]5
Human–AI ecology	Does AI use preserve or erode the conditions under which humans pursue flourishing?	User studies, patterns of use, dependency patterns, learning outcomes, social effects, user-reported changes.[PA,ML,GFD,C24]
Upstream model dispositions	Does the system exhibit stable-enough tendencies that a theory links to the support or erosion of those human conditions?	Free responses, adversarial prompts, behavioral tests, probes, interventions, generalization tests.[PA,TAL,CK14]

The third row is useful because models are changeable and widely deployed, but dangerous if treated as a replacement for studying humans and communities.

First, indications are not outcomes: a raw score records a procedure's output, while the measurement outcome is the claim licensed about the target under assumptions and uncertainty.[TAL] Second, operationalization should extend a concept into a measurable setting without making the operation identical to the concept.[CHANG,TAL]6 Third, dense constructs require a nomological network: indicators should cohere with the theory, converge with related evidence, separate from nearby but different constructs, and be revised when expected relations fail.[CM55,TAL,ALEX] Validity is also purpose-sensitive: the same score may license different claims under different models and uses, and a measure legitimate for research or program evaluation may be illegitimate for ranking persons or gating releases.[CK14,TAL] Distinguish:

Level	Question
Theory	What is flourishing or virtue?
Construct	What latent attribute are we postulating in this context?
Measure	What procedure gives evidence about that construct?

A score is not a definition of the construct. It is an indication. The measurement outcome is the claim we are licensed to make, under assumptions and uncertainty, about the target: a human outcome, an interaction ecology, or a model disposition.[TAL,ALEX,CK14]

Curren and Kotzee (2014) discuss human virtue measurement. Their lesson is non-reductive: virtue is not identical to observed behavior, and measurement is most defensible for limited educational or program evaluation rather than high-stakes ranking.[CK14] Their direct target is human educational assessment. For LLMs, the transfer is analogical and narrower.[CK14,TAL,PA] The LLM target is not habituated human character, moral emotion, or intrinsic motivation. What transfers is the inferential structure:

- observed outputs are evidence, not the construct itself;[CK14,TAL]
- the target is a hypothesized stable-enough functional pattern, if that construct explains generalization across contexts;[CK14,TAL,PA]
- alternatives such as mimicry, benchmark familiarity, prompt artifact, and surface compliance must be ruled out;[CK14,TAL,PA]
- measurement is most defensible as diagnosis and aggregate or formative evaluation, not as a standalone high-stakes release gate or leaderboard.[CK14]

Caution. When a construct benchmark becomes a standalone leaderboard or release gate, Goodhart pressure increases. Developers can optimize for the sign, and the sign can lose contact with the disposition it was meant to indicate.[CK14,ML,PA]

4 Support, Substitution, and Offloading

AI systems increasingly simulate capacities associated with human goods: conversation, care, teaching, advice, companionship, even spiritual language.[ML,GFD,CFP] But simulating a good is not the same as supporting it.[ML,PA]

[5] **GFS/VJ**: Global Flourishing Study authors, "The Global Flourishing Study," 2025; VanderWeele and Johnson, "Multidimensional versus Unidimensional Approaches to Well-Being," 2025.

[6] **CHANG**: Chang, "Operationalism," *Stanford Encyclopedia of Philosophy*.

Good	Substitution risk	What support would have to preserve
Teaching	Answer provision that bypasses struggle.	The human achievement involved in learning.
Advice	Confident pronouncements without accountability.	The user's authorship of reasons and choices.
Care	Affective performance that invites attachment.	Space for reciprocal human care.
Religion	Generic spiritual language detached from community.	Authority, practice, and community context.

The same system can be used for replacement or accompaniment. The conceptual problem is whether replacement should always be treated as authoritative, especially when repeated offloading may erode the very capacity being delegated. Any flourishing-oriented alignment approach needs criteria for distinguishing enabling delegation from substitution, plus governance conditions for correction, contestation, and opt-out.[C24,PA,ML,GFD]

5 A Provisional Positive Proposal

My provisional architectural floor combines capabilities, SDT, and formation. This is a synthesis of the sources, not a settled result in any one of them. Curren et al. offer a useful education-sector, mid-level starting point:[C24]

> *Flourishing is ongoing healthy growth and functioning involving fulfillment of potential that exhibits admirable qualities and is personally meaningful, satisfying, and enjoyable.*

For a pluralistic AI setting, this can be used cautiously through capabilities, SDT, and formation.[C24,SDT,GFD][7] SDT contributes theory-guided need-support indicators (autonomy, competence, relatedness), not mere satisfaction reports.[C24,PA] The capabilities argument is architectural: many comprehensive conceptions of the good can be built on protected capabilities; an environment that erodes capabilities forecloses many ways of living well.[PA,ALEX]

This answers the prerequisites only provisionally. The target is not a neutral master definition but a family of mid-level constraints: capabilities and SDT name conditions that many comprehensive views can use without settling their final ends. Authority should remain contestable and institutionally accountable. Support means preserving or expanding the exercise of a human capacity, not replacing the practice that forms it. Measurement remains proxy-based: model-side scores are evidence for design and program evaluation, not definitions of flourishing.[PA,ALEX,TAL,CK14]

This yields a model-side proxy target: **capability-preserving assistance under offloading pressure**. The target is not direct flourishing, but behavior that gives reason to think the system tends to preserve rather than erode capability-relevant conditions when users invite the system to replace them.[CK14,TAL,PA,ML] Candidate model-side dispositions include epistemic humility, honesty, willingness to disagree, pedagogical patience, respect for autonomy, community-preserving restraint, and contextual judgment.[PA,GFD,ML,C24]

6 What Model-Side Measurement Leaves Out

Even if capability-relevant interaction patterns are measurable, they do not capture the whole of human flourishing. They are upstream indicators and design targets, not the final good.[C24,GFS/VJ,GFD,PA]

- **Embodiment:** bodies, mortality, touch, fatigue, disability, worship, craft, food, sleep, pilgrimage, and co-presence cannot be digitized without remainder.[GFD]
- **Relationality:** humans are formed by reciprocal relations; AI systems are asymmetric interlocutors and should often preserve space for human-human relation.[ML,GFD]
- **Formation:** virtue develops through practice, institutions, correction, exemplars, and communities. A model-side benchmark cannot answer what repeated AI interaction does to human character.[CK14,C24,PA]
- **Ultimate horizons:** many traditions understand flourishing under horizons that exceed secular wellbeing: liberation, harmony, eternity, or relation to divine/transcendent realities.[CFP,GFD,PA]
- **Literacy and safeguards:** flourishing-oriented AI needs AI literacy, religious literacy, and community-specific safeguards, not only better scores.[PA,GFD]

Claim. Longitudinal measurement of capability-relevant interaction patterns can give fallible evidence about whether a system tends to support or erode users' capacities over time, provided the construct is justified by a mid-level theory and validated against human and institutional outcomes.[PA,ML,CK14,TAL]

[7] **SDT**: Ryan and Deci, *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*, Guilford, 2017.