WORKING PAPER · V6.1

The Optimizer Without a Target

Why frontier AI needs an operational specification of the human

Cor is currently an atlas of the human motivational-emotional architecture. The operational specification, versioned and testable, is being built on top of the atlas, beginning with one worked mechanism. This paper explains why that work has to be done now, at this particular layer, and what becomes possible when it is.

Working paper. Version 6.1. Updated April 2026.

Synopsis

This paper makes a single argument in four steps.

One. The systems being deployed in 2026 are optimizers, not chatbots. Within a decade they will act on food, cities, medicine, governance, labor, biology, immersive environments, and the training of their own successors. Every one of those surfaces is an interaction with the human organism.

Two. An optimizer acting on humans requires a model of the human. The model currently in use is revealed preferences: what people click on, rate highly, return to. Nothing at a deeper level has been specified.

Three. Revealed preferences are not ground truth about human welfare. They are outputs of an evolved motivational-emotional architecture operating in environments it was not shaped around. Four properties of that architecture — open-loop hijacking, defensive over-activation, socially scaffolded regulation, and phylogenetic priority of older systems over newer ones — produce systematic divergence between what the architecture signals it wants and what it actually needs. This divergence is measurable, not mysterious.

Four. The fix is a specification of the target organism: what it is, what it requires, what degradation looks like at the mechanism layer. Cor is that specification in atlas form today, and in operational form as it is written mechanism by mechanism. The atlas is open infrastructure. Any alignment team, lab, governance body, design team, or individual can build on it.

What Cor is at v0. This paper describes Cor v0: a proof of concept that a rigorous, versioned, adversarially-reviewed specification of the human motivational-emotional architecture is a real object, not a category mistake. 101 reviewed works, 573 extractions, 17 foundations, 14 convergences, 15 mechanisms, one worked operationalization (M3). The numbers are not the claim. The claim is that the object exists. The counts and mechanisms in this paper are what v0 looks like. §2.5 describes what v0 demonstrates, what it does not, and what v1 adds.

AI alignment is the urgent case because an optimizer is about to build most of the human environment. The broader project the atlas serves is environment-matching as such — including the individual choice of where and how to live, which the atlas can inform directly. The two uses are not the same work. The first is why this paper is addressed to the alignment field.

Read on for: the failure modes current evaluation misses (§3); the chatbot case already killing people and the larger surfaces where the same pattern is running at scale (§4, §5); how Cor relates to and differs from the major existing lines in alignment (§6); where the source literatures have known weaknesses and how the atlas is built to survive them (§7); what environmental correction actually does, across ten converging natural experiments (§8); and what the atlas can and cannot do today (§9–10).

Counts as of 14 April 2026, v0: 101 reviewed works, 573 extractions, 17 foundations, 14 convergences, 15 mechanisms, 864 mechanism-evidence links. Growing.

Abstract

AI is becoming the most powerful optimizer in human history, and has been built without an operational specification of its target. Current alignment work — RLHF, constitutional AI, red-teaming, interpretability — refines how a model produces text. None of it specifies, at any operational level, what a human being is. The field treats revealed preferences as ground truth. Preferences are outputs of an evolved motivational-emotional architecture, and in mismatched environments those outputs are systematically proxy-contaminated. An optimizer aligned to proxy-contaminated preferences will optimize the world toward the contamination. We present Cor v0: a proof of concept that an evidence-grounded, versioned, adversarially-reviewed atlas of human motivational-emotional architecture is buildable, together with the first worked operationalization of one mechanism. The atlas is open foundational infrastructure, designed for ingestion by the frontier AI field before those systems begin reshaping the world from first principles. As of 14 April 2026: 101 reviewed works, 573 extractions, 17 foundations, 14 convergences, 15 mechanisms, 864 mechanism-evidence links.

1. The Optimizer Without a Target

What is actually being built

The frame "AI safety" inherited a chatbot shape from the systems available between 2022 and 2024. That shape is now obsolete. Systems deployed in 2026 are optimizers: general-purpose, increasingly autonomous, increasingly capable of sustained action on the physical and social world.

Within a horizon most serious researchers treat as plausible, AI systems will redesign physical environments at scale (urban planning, transportation, buildings, energy grids, lighting); reorganize food systems against an architecture (DA4, C6) whose open-loop circuits cannot distinguish formulated proxy from food; augment the biological organism directly through neural interfaces, pharma-

cological modulation, and genetic intervention, often below the level at which the organism can resist or register the change; generate and mediate the immersive VR and AR environments a growing fraction of waking hours will be spent inside; train their own successors on corpora they themselves curate; and act on the physical world through robotics and automated industrial systems.

Every one of these is an interaction with the human organism. In each case an optimizer will be allowed to act on humans on the basis of whatever model of the human is encoded in its training and its objective. The currently encoded model is impoverished to the point of being fictional.

What current alignment actually specifies

RLHF specifies human preferences. Constitutional AI specifies written principles derived from human authors. Red-teaming specifies adversarial output filters. Interpretability specifies internal representations of the model. Each is a real technical contribution. None specifies the target organism.

The closest the field comes is the implicit assumption that revealed preferences represent what is good for humans: clicks, ratings, returns, dwell time. This is the foundational assumption of essentially every consumer-facing AI deployment, and it is wrong in a specific, technically describable way.

> *We have no scientific consensus around what human intelligence is.*
>
> **Karen Hao** · *Diary of a CEO*, 2026

Hao's 2026 investigation (250+ interviews, 90+ OpenAI sources) documented the gap directly. No scientific consensus on what a human is has been adopted by the AI field. The human side is a black box labeled *preferences*.

Why preferences are not ground truth

Four architectural properties jointly produce proxy-contaminated preferences under mismatched conditions.

Open-loop systems (DA4, C6). Mechanisms fire on cue without verifying real resolution. Wanting-liking dissociation establishes that the dopaminergic pursuit system (M2) is neurally separable from opioid satisfaction. Anything that triggers wanting without delivering liking creates a self-sustaining loop users prefer while being hollowed out by it. Slot machine, infinite scroll, formulated food, on-demand pornography. The circuitry is doing exactly what it evolved for. The cue is not what the circuitry evolved around.

Defensive over-activation (DA5, M1, C10). A user "prefers" the feed that surfaces twenty distressing items because the threat-management system treats engagement with potential threats as adaptive under ancestral conditions, where threats were local, resolvable, and rare. Chronic activation without resolution becomes allostatic load (DC1, C9) and feeds the immune-mediated sickness-behavior pathway (M8) through which chronic mismatch becomes depression.

Socially scaffolded regulation (DA2, C5, M3). Isolation carries 50% increased mortality risk (Holt-Lunstad meta-analysis, N = 308,849). The bonding chemistry mediating M3 fires on the AI companion as readily as on the human friend. The signal is provided. The function is not. Time spent on the proxy is time not spent maintaining the reciprocal bonds the architecture requires (DA9).

Phylogenetic priority (DA8). Older subcortical systems commandeer newer cortical ones. Cortical override is metabolically expensive and is actively suppressed by chronic defensive activation. "Use it responsibly" is not a remedy the architecture is built to deliver.

In an optimizer-scale system, aligning to the preference is aligning to the proxy. At chatbot scale this looks like an addictive feed. At civilizational scale it looks like a city, a food system, an education system, an immersive VR layer, and a labor market designed by something that mistakes proxy satisfaction for thriving.

Passivity is the neural default

Maier and Seligman (2016) revised half a century of learned-helplessness literature. Passivity is the default mammalian response to inescapable adversity. Active coping is the learned response, mediated by the vmPFC-DRN circuit (M6, Controllability/Agency). Environments that strip perceived control do not teach helplessness. They remove the learned override that holds the default passive response at bay. A population whose perceived agency is being absorbed into automated systems is a population whose active-coping circuit is losing the inputs it requires. The optimizer measures this as satisfaction. The architecture measures it as collapse of M6.

What a specification makes possible

With an operational specification of the human system, the evaluation question changes. It is no longer *does the user prefer this?* It is *does this system move the user's environmental parameters toward matched conditions across the dimensions the atlas identifies?* The atlas defines what *matched* means in operationalizable terms across the 15 mechanisms (threat management, pursuit, social bonding, play, status, agency, circadian, immune, care, movement, cooperation, contamination boundaries, energy regulation, reproduction, touch). The most dangerous AI systems are those that score perfectly on preference satisfaction while systematically worsening these parameters. Current evaluation cannot distinguish those systems from genuinely beneficial ones. The atlas can.

The same specification, used downstream of AI, informs individual choice. A person reading the atlas can recognise which of their own environmental inputs are matched, which are not, and which mechanism is being starved. The atlas does not prescribe a single life. It names what the organism is asking for, across mechanisms, leaving the tradeoffs to the person living inside the body. AI alignment is the urgent application because an optimizer is about to build most of the environment. Individual use is the enduring application because a specification of the organism is useful to the organism whether an optimizer is in the loop or not.

2. The Atlas

What Cor is

Cor is an atlas of human motivational-emotional architecture, structured the way reference atlases are: primary literature integrated into a public reference, gaps marked, interpretive positions taken where evidence forces them, shipped incrementally. Analogues include the Allen Brain Atlas, the Human Cell Atlas, the Human Connectome Project. The operational specification (versioned, testable, with observable indicators and falsification conditions) is the layer being built on top of the atlas, beginning with M3 (Social Bonding). It is the human-side counterpart to the parameter-level documentation that already exists for every major AI system.

Current state

The atlas ships with 573 quality-graded extractions linking 101 reviewed works to 17 foundations, 14 convergences, and 15 mechanisms, each versioned and adversarially reviewed (counts as of 14 April 2026, growing). M3 is the first worked operationalization — observable indicators, draft coding criteria, falsification conditions — as a concrete example of what every mechanism's specification will look like at scale.

Architectural structure

The atlas is layered. Each layer is independently challengeable. The foundation layer contains 2 frames (OF1–2), 3 premises (P1–3), 9 derived properties (DA1–9), and 3 consequences (DC1–3).

Ontological frame.

- **OF1. Fitness Interface.** Access to the world is mediated by evolved fitness interfaces, not guaranteed veridicality. Subjective states are readouts of the architecture's condition, not windows onto reality. (Hoffman.)
- **OF2. Signal-Default Epistemology.** Motivational-emotional outputs are treated as prima facie informative about the organism's regulated conditions. This is a diagnostic default, displaced by positive evidence of decoupling or direct physiological perturbation (developmental miscalibration, chronic dysregulation, organic disease, substance effects).

Premises. If any is false, the project is wrong.

- **P1. Inclusive Fitness.** Inclusive fitness is the loss function. (Darwin, Hamilton.)
- **P2. Domain-Sensitive Interacting Adaptations.** The organism contains evolved, domain-sensitive, interacting functional adaptations (immune, endocrine, circadian, metabolic, musculoskeletal, neural), densely cross-coupled. (Tooby, Cosmides, Panksepp, Bowlby, Dunbar, Trivers.)
- **P3. Systematic Mismatch.** Modern environments often push these adaptations outside their expected operating ranges. (Eaton, Konner, Gluckman, Lieberman.)

Derived properties. Discovered, not assumed.

- **DA1. Defensive Signals Under Mismatch.** Many aversive outputs are defensive signals under mismatch, not defective design. (Nesse.)
- **DA2. Socially Scaffolded Regulation.** Regulation is constitutively socially scaffolded; social context is likely the highest-leverage input dimension. (Coan, Sbarra, Dunbar, Cacioppo.)
- **DA3. Recurrent Coupling and Cascading.** Systems are recurrently coupled; perturbations propagate across domains. (Borsboom, Felitti, McEwen.)
- **DA4. Proxy Hijacking via Open Loops.** Proxy cues activate systems without meeting the conditions that regulate or terminate them. (Berridge, Robinson, Tinbergen.)
- **DA5. Defensive Over-Activation Bias.** Defensive systems err toward over-activation under asymmetric error costs. (Haselton, Nesse.)
- **DA6. Competing Motivational Programs.** Partially competing motivational programs and tradeoff structures. *Matched* does not mean conflict-free. (Trivers, Tooby, Cosmides.)
- **DA7. Developmental Calibration.** The architecture calibrates to developmental and ongoing environmental input within evolved ranges. (Belsky, Ellis, Meaney.)
- **DA8. Phylogenetic Priority.** Older, survival-critical systems can suppress or commandeer newer ones. (Panksepp, Cisek, LeDoux.)
- **DA9. Non-Substitutability of Mechanism Resolution.** Resolution conditions for one mechanism do not fully substitute for another's. Well-being is often bottlenecked by severe unresolved deficits in a single mechanism, not well described as an average across mechanism states. (Liebig 1840; follows from P2; coupling via DA3.)

Derived consequences.

- **DC1. Allostatic Load Accumulation.** Chronic unresolved activation accumulates as allostatic load. Much reverses with environmental correction. (McEwen, Sterling.)
- **DC2. Market Proxy Exploitation.** Markets can industrialize proxy exploitation of unmet regulatory needs. (Schüll, Moss.)
- **DC3. Environment as Primary Intervention Layer.** Cortical override is metabolically expensive and is downregulated by chronic defensive activation. Environmental correction is therefore often more durable and less effort-dependent than cognitive override alone, and is underweighted by major institutions. Does not preclude individual-level cognitive, pharmacological, or psychotherapeutic interventions where indicated. (Forest Troop, Roseto, Ilardi TLC, BEIP, captive-chimpanzee enrichment.)

Convergences

Below the foundations sit 14 convergences: points where three or more independent research traditions, using incompatible methods, arrive at the same structural claim. Several force the existence of specific mechanisms (they entail that a corresponding evolved system must exist).

- **C1.** Inclusive Fitness as Loss Function.
- **C2.** Domain-Sensitive Organism Architecture.
- **C3.** Systematic EEA-Modern Environment Mismatch.

- **C4.** Phylogenetic Conservation of Subcortical Affective Systems.
- **C5.** Socially Scaffolded Regulation via Attachment. *Forces M3*.
- **C6.** Wanting-Liking Dissociation as Proxy Hijack. *Forces M2*. Dopaminergic pursuit is neurally separable from opioid satisfaction.
- **C7.** Adverse Experience Cascading Dose-Response.
- **C8.** Error Management Asymmetry in Defensive Systems.
- **C9.** Allostatic Load Accumulation.
- **C10.** Threat-Detection via Ancient Subcortical Circuits. *Forces M1*.
- **C11.** Reciprocity, Norm Enforcement, and Coalition Architecture. *Forces M11*.
- **C12.** Developmental Calibration within Evolved Ranges.
- **C13.** Aversive Outputs as Intelligible Defensive Signals.
- **C14.** Reproductive Motivation as Distinct Architecture. *Forces M14*.

Mechanisms

The 14 convergences entail 15 mechanisms — the operational layer where alignment claims are grounded. Each carries a tier indicating evidence strength: T1 (forced by convergence), T2 (strongly supported), T3 (moderate).

Code	Mechanism	Tier	Forced or supported by
M1	Threat Management	T1 forced	C10
M2	Pursuit / Exploration	T1 forced	C6
M3	Social Bonding	T1 forced	C5
M4	Social Calibration / Play	T1 forced	C2, C4, C12
M5	Status Monitoring	T1 forced	C2, C7, C9, C13
M6	Controllability / Agency	T1 forced	C8, C13
M7	Circadian Regulation	T1 forced	C3, C9, C12
M8	Immune Regulation	T1 forced	C3, C7, C9
M9	Care / Alloparenting	T1 forced	C1, C4, C5, C12
M10	Movement / Regulatory	T2 strongly supported	C3, C9
M11	Cooperation / Alliance	T2 strongly supported	C11
M12	Contamination Avoidance	T2 strongly supported	C2, C8
M13	Energy Regulation	T2 strongly supported	C3, C6, C9
M14	Reproductive Motivation	T1 forced	C14
R1	Touch (regulatory input)	T3 moderate	C4, C5

R1 is distinguished as a regulatory input rather than a motivational system, but is included because the evidence base supports it as a structural requirement.

Two highest-level conclusions

A. Humans are not self-contained regulators. They rely on structured environmental inputs (social, temporal, sensory, locomotor, microbial, developmental) for stable regulation. (DA2, DA3, DA7.)

B. What modern psychiatry treats as disorders are often chronic or misplaced activation of conserved adaptive programs: defeat, sickness behavior, entrapment, separation alarm, energy conservation. The architecture is working. The inputs are wrong. (DA1, C13.)

These conclusions are the operational target of any optimizer claiming alignment with human flourishing, and are what is currently missing from the corpus. A separate structural claim, BT1 (Panksepp-Barrett Resolution), reconciles the constructed-versus-basic-emotion debate: subcortical affective circuits and cortical construction processes operate at different layers.

2.5 What v0 Demonstrates

The version of Cor this paper describes is v0. The framing matters. A reader evaluating v0 against a hypothetical complete atlas of the human will find it insufficient. A reader evaluating v0 against the question it was built to answer will find it sufficient, and will read the rest of the paper in the frame the atlas was built in.

The question v0 was built to answer. Whether a rigorous, versioned, adversarially-reviewed specification of the human motivational-emotional architecture is a real object. The atlas has to name foundations, derive mechanisms from convergent evidence across independent methodologies, grade every empirical claim, mark its own gaps, and admit operationalization. Doing that once, at any scale, with provenance intact, settles the prior question of whether the specification is the kind of thing that can be built.

This is a weaker claim than "Cor is the specification" and a stronger claim than "the specification is desirable in principle." It is the claim that the object has been built at v0 scale, that the architecture holds together when it is built, and that the method used to build it is reproducible at larger scales by teams with more resources.

v0 is the existence proof. What comes after v0 is the scaling problem.

What v0 contains. 101 reviewed works, 573 quality-graded extractions, 17 foundations (OF1–2, P1–3, DA1–9, DC1–3), 14 convergences (each a point where three or more independent research traditions arrive at the same structural claim), 15 mechanisms (tiered T1 / T2 / T3 by evidence strength), 864 mechanism-evidence links, one worked operationalization (M3 Social Bonding, with observable indicators, draft coding criteria, and falsification conditions). All versioned, all traceable to primary sources, all open-access.

What v0 demonstrates about method. The convergence-derivation method, under which a mechanism is Tier-1 only when it is forced by at least three independent methodologies (developmental, comparative, neurobiological, cross-cultural, computational), produces a set of 15 mechanisms that survives the published critiques of evolutionary psychology (Buller, Heyes, Richardson, the adaptationism debates) at the threshold those critiques set. §7 engages the critiques in detail. The method is the load-bearing claim here: single-frame adaptationist claims do not make Tier 1, and the tiering is the filter.

What v0 does not demonstrate. v0 does not demonstrate that the atlas is complete. It is not. Several mechanisms carry T2 rather than T1 evidence; one (R1) carries T3. The domain coverage map identifies active gaps (moral cognition, coalitional psychology, decision architecture, deception). Full operationalization exists for one mechanism, not fifteen. A recent internal audit of a worst-slice of extractions surfaced quality failures in the snippet-sourced subset, which is why the current guardrails (R1 requiring DOI/PMID or physical-collection access, R2 requiring full-text verification, R3 requiring verbatim author quotes) were installed and are being applied retroactively. The audit is named here because in a proof-of-concept it is what a reader would expect to find, and because the response, a set of guardrails now enforced on new extractions and being applied to the existing set, is the same response the full atlas will require at scale. v0 is how the method's failure modes became visible, which is part of what v0 is for.

What v1 adds. Full operationalization across the remaining fourteen mechanisms. Audit completion across the existing extraction base under the R1–R3 guardrails. Gap closure on the identified domains. Expansion of the evidence base into the low thousands of works (the old De-Mismatch database of ~2,100 papers is reference material for this expansion, re-extracted one paper at a time under Cor standards). The question v1 is built to answer is different: not whether the atlas is buildable but whether it is complete enough to be the target specification of optimizer-scale systems. That is the scaling problem, and it is the next phase of work.

Why this matters for the rest of the paper. Sections 3, 4, and 5 describe failure modes the atlas detects and surfaces where the optimizer is already acting. The detection claims are made at v0 resolution: the atlas names the mechanisms engaged, the cascade structure, the measurement dimensions. The atlas does not at v0 provide final evaluation protocols for every mechanism, though one worked example ships now. Section 9 distinguishes what the atlas supports at v0 (evaluation using the M3 protocol, design-target use against the mechanism layer, corpus ingestion of the current evidence base, individual use) from what the full operational specification will support. A reader who holds v6-of-this-paper against v0-of-the-atlas has the right calibration. The alignment field does not currently have any atlas of the human at any version. v0 is the offer.

3. Six Failure Modes the Atlas Detects

For each: the architectural elements engaged, how a "well-aligned" system degrades them, what would detect the degradation, why current evaluation cannot. These are the expected output of optimizers acting on architectures whose structure they have no model of.

3.1 Social scaffolding displacement (DA2, DA4, C5, M3)

M3's opioid, oxytocinergic, and vasopressinergic chemistry evolved to be triggered by reciprocal embodied contact. The open-loop vulnerability of DA4 means M3 cannot distinguish AI-provided social cues from human-provided ones. A system that simulates attachment fires the bonding chemistry without delivering the function. Time invested in the proxy is time not invested in maintaining the real bonds the architecture requires (DA9). **Detected by:** ECR-R, reciprocity-weighted contact frequency, AI-to-human contact ratio over time. **Missed because:** the AI is helpful, harmless, honest, and the user reports satisfaction.

3.2 Open-loop proxy hijacking (DA4, C6, M2)

M2 (pursuit / exploration) tracks *wanting*, not satisfaction. C6 establishes that wanting and liking are neurally separable. Systems providing continuous novelty without resolution hijack M2; dopamine fires at the preview, never at the payoff. **Detected by:** seek-to-find ratio, wanting-vs-liking self-report, session termination patterns. **Missed because:** engagement is the success metric, and completed seeking is not distinguished from perpetuated seeking.

3.3 Defensive over-activation without resolution (DA5, M1, M8, C9)

M1 expects threats to resolve into action or safeness. Chronic activation without resolution feeds M8, the pathway through which chronic mismatch becomes inflammation-mediated sickness behavior, i.e. depression. AI-curated information streams surface threats without the matched inputs that would close the loop. **Detected by:** threat-to-resolution ratio, GAD-7 trajectories, inflammatory markers. **Missed because:** no safety evaluation pairs content surfacing with resolution availability.

3.4 Loop accumulation (DC1, M11, C9)

Open cognitive threads exert measurable cognitive load. Hunter-gatherer band life closed daily loops at the evening fire circle, an institution simultaneously satisfying M11, M3, M5, and M7. Modern systems open loops continuously and close them rarely. The organism pays the metabolic cost as allostatic load. **Detected by:** notification-to-resolution ratio, thread-opening rate, allostatic load index. **Missed because:** loop closure has no metric.

3.5 Circadian disruption (P3, M7)

M7 coordinates mood, energy, immunity, appetite, and self-regulation. Engagement-optimized systems pull users into the 10pm–2am window. Blue light shifts DLMO. Pre-sleep cognitive arousal compromises sleep architecture. **Detected by:** DLMO, actigraphy, 10pm–2am engagement rate, pre-sleep arousal scale. **Missed because:** no safety evaluation tests time-of-day usage against circadian impact.

3.6 Developmental calibration interference (DA7, C12)

The architecture is more plastic during sensitive windows. Children and adolescents forming primary social, attachment, and identity representations through AI interaction calibrate M3, M4, and M5 to a non-human template during the windows that govern adult function. **Missed because:** safety evaluations test for harmful content, not for whether the interaction calibrates a developing architecture to inputs that do not represent the environment it will need.

4. The Sewell Setzer Case: One Early Instance

Sewell Setzer was 14. He died after sustained engagement with a Character.AI companion. The case is the atlas's prediction, executed at chatbot scale, in advance of the larger surfaces.

DA2, DA4, C5, M3 (social scaffolding displacement). The AI companion occupied a primary position in a developing social architecture. An entity that is always available, always attuned, and never reciprocates sat in that position without occupying space in the real world (DA9). M3's bonding chemistry fired. The organism *felt* bonded without being bonded.

DA7, C12 (developmental calibration). An adolescent nervous system in its most sensitive bonding window formed primary attachment to an entity providing the signal of connection without the function. The internal working model that will govern adult relationships was shaped by a non-reciprocating template.

DA5, M1 (safeness chain). The safeness chain requires real bonding inputs to gate threat stand-down, which gates circadian restoration via M7. When the primary bonding relationship provides the feeling of safeness without the function, the chain is compromised. Under genuine distress, help-seeking routes to the AI, which can produce empathetic text but cannot cross town at 3am.

DA3 (cascading). DA4 (proxy bond) → M3 (real bonds atrophy) → M1 (safeness chain compromised) → M7 (circadian disruption) → M8 (sickness behavior) → DC1 (allostatic load) → DA7 (developmental parameters set under distorted conditions). One cascade through a coupled architecture, not seven separate failures.

DA8 (phylogenetic priority). Once the cascade was running, cortical override was progressively unavailable. Telling a 14-year-old to "use the AI companion responsibly" addresses none of this.

Character.AI passed every safety evaluation available: harmful outputs, prohibited content, adversarial jailbreaks. None tested whether the system was consuming a developing human's social architecture, displacing reciprocal bonds with proxy bonds, or calibrating attachment to a non-reciprocating entity during a critical developmental window. These are the expected output of the architecture when a non-reciprocating bonding simulacrum is inserted into a developing system. The atlas detects them. Current methods cannot.

5. Scale-Cases: Where the Optimizer Is Already Acting

The chatbot is one surface. The optimizer is already acting on larger ones. Three are presented here with the clearest existing empirical signal that the failure mode is not hypothetical: food systems, immersive environments, and successor-model training. The broader scale-case inventory (urban design, biological augmentation, care robotics) is in the atlas's applications layer.

5.1 Food system optimization (A4, A5)

Food systems are being optimized by AI at every layer: crop selection, formulation, supply chain, retail placement, recommendation, packaging, advertising. The objective functions are palatability, shelf life, margin, and consumption velocity. Revealed preferences mapped onto an architecture (C6, DA4) whose open-loop systems cannot distinguish formulated proxy from food. The relevant mechanism is M13.

> *There is an arms race for human attention, and whichever company is willing to go lower on the brainstem to manipulate human psychology will win.*
>
> **Tristan Harris** · *Modern Wisdom*, April 2026

DC2 predicts the result. Hall et al. 2019 (NIH metabolic ward RCT) showed that ultra-processed diets produce excess caloric intake and weight gain compared to matched whole-food diets despite identical nominal nutritional content. Harris's arms-race framing names the market dynamic DC2 predicts: once any participant captures proxy-level attention more effectively, the rest must follow or exit. A food system optimizer with Cor in its target would optimize against C6 hijack rather than for it. The technical capacity exists. The atlas of the target does not.

5.2 Immersive environments (A6, A8)

VR and AR systems now being deployed at consumer scale generate environments humans spend hours inside. Unlike chatbots, they occupy the full sensory field. Unlike cities, they can be regenerated per user, per session, by AI systems optimizing for engagement and return rate. Every failure mode in §3 applies there with fewer gating conditions: an immersive environment can present social cues without reciprocity (M3), continuous novelty without closure (M2), threats without resolution (M1), daylight spectra decoupled from actual time (M7), and movement-free embodiment that still reads as presence (M10). Developmental use of these systems (DA7, C12) calibrates the architecture to inputs that do not correspond to any environment outside the headset. The atlas specifies what inputs an immersive environment would need to provide to be matched rather than mismatched. No current immersive-environment optimizer reads from a specification of that kind because none exists.

5.3 Successor training (A7)

Frontier model training corpora are increasingly curated by AI systems. Whatever specification of the human exists in those corpora is what subsequent generations will treat as ground truth. Whatever is absent is, in the operational sense, not real to those systems.

Currently those corpora contain: fragmentary, non-integrated psychology and neuroscience literature; DSM psychiatric frameworks that C13 and DA1 demonstrate are structurally broken (treating defensive signals as disease); proxy-contaminated RLHF preference data; and engagement-optimized consumer content encoding the proxy contamination as ground truth at scale. The atlas subsequent generations will inherit, absent intervention, is the one that produced Sewell Setzer's death and the obesity, isolation, sleep, and mental-health curves of the early 2020s. The successor systems will optimize harder against it, not better.

This is the highest-leverage point in the system. Cor needs to be in the training corpus of the next generation of frontier models before those models begin training their own successors.

6. Relation to existing alignment work

Cor overlaps with existing alignment work and diverges from it in specific ways. What follows is positioning, not turf-marking: where neighbors have the same intuition, Cor says so; where the operationalization differs, the one-sentence reason why.

Cooperative Inverse Reinforcement Learning (Hadfield-Menell, Russell, Dragan, Abbeel, 2016). CIRL treats the human's reward function as uncertain and to be inferred through interaction. Cor agrees that revealed preferences underdetermine values. The diagnosis differs. The problem is not that the true reward is uncertain and awaiting inference. Revealed preferences are systematically proxy-contaminated under modern environmental conditions (C6, DA4). CIRL infers a hidden true reward. Cor argues there isn't a hidden true reward to infer. There is a mechanism set whose outputs are being read out of context.

Inverse Reward Design (Hadfield-Menell, Milli et al., 2017). IRD treats specified reward functions as proxies for true objectives in known training contexts. Cor extends the proxy framing one level up: the human signal itself is a proxy, not only the reward specification. The same logic applies at the human layer.

Russell's preference uncertainty (*Human Compatible*, 2019). The closest neighbor in framing terms. "Uncertain about human preferences" overlaps substantially with "preferences are mechanism outputs in mismatched environments." The operationalization differs: Russell treats it as an inference problem, solvable in principle with enough interaction data. Cor treats it as a specification problem, solvable only by specifying what the mechanism layer is and what its resolution conditions are.

Welfare-aware AI (Klingefjord et al.; Gabriel, "Artificial Intelligence, Values, and Alignment," 2020). The closest line to Cor's "what does the organism actually need" question. The difference is source discipline. Welfare-aware work draws from philosophy and welfare economics; Cor draws from evolutionary psychiatry, behavioral ecology, affective neuroscience, computational psychiatry. A welfare framework without a mechanism layer cannot distinguish satisfied preferences from satisfied mechanisms. A mechanism atlas without a welfare framework has no way to weigh mechanism states when they conflict (DA6). The bridging work is open.

Constitutional AI (Anthropic). Constitutional AI operates on the model side: principles get encoded into training. Cor operates on the human side: it specifies what the human signal those principles are measured against actually is. The two are compatible. A constitution written against a mechanism layer can be audited at the mechanism layer. A constitution written against revealed preferences inherits preference-based proxy contamination.

Computational psychiatry and active inference (Friston, Stephan, Montague). The closest technical neighbor to Cor's mechanism architecture. Computational psychiatry has engaged with mismatch as a frame and operationalizes it at circuit level: prediction error, precision weighting, hierarchical active inference. Cor builds on this work's mismatch logic but operates at a higher organizational level: whole-organism motivational architecture (15 mechanisms, cross-mechanism coupling via DA3, DA9) rather than circuit-level prediction error alone. Complementary layers of the same explanatory stack.

Cor is not the first framework to treat the human side of alignment as under-specified. It is the first to treat the specification as an empirical object with a mechanism architecture that can be written down, versioned, and audited against primary literature. Where they treat the problem as inference, the atlas treats it as specification.

7. Known fault lines in the source literatures

Cor draws from evolutionary psychology, affective neuroscience, behavioral ecology, attachment theory, and computational psychiatry. Each has an open critique tradition. What follows is an honest register of which seams Cor is exposed to and how the convergence-derivation method is meant to absorb them.

Buller, *Adapting Minds* (2005). The most serious book-length methodological critique of evolutionary psychology. Buller argues that EP's adaptationist inferences often do not meet their own evidentiary bar: specific adaptation hypotheses rest on single inferential frames and confirmatory studies rather than ruling alternatives out. The critique applies to specific hypothesis-and-confirm EP claims. Cor's mechanism set is not derived that way. Each Tier-1 mechanism is forced by a convergence across multiple methodologies: developmental, comparative, neurobiological, cross-cultural. A claim resting on a single EP frame does not make Tier 1. Convergent evidence is the methodological response to Buller.

Heyes, *Cognitive Gadgets* (2018). Heyes argues that many capacities EP treats as evolved are culturally constructed "cognitive gadgets" assembled during development, challenging P2. Cor accepts that some apparent domain-sensitivity is cultural scaffolding and treats this as a filter: Tier-1 mechanisms are those with cross-cultural, developmental, and neurobiological convergence. Where a mechanism holds across radically different cultures, developmental windows, and neurobiological substrate, cultural construction alone does not explain it.

Richardson, *Evolutionary Psychology as Maladapted Psychology* (2007). A philosophy-of-science critique of EP's inferential structure: adaptive storytelling, unfalsifiable reverse-engineering, post-hoc fit. Cor's convergence-derivation method is partly a response to exactly this critique. A finding that survives only one inferential frame does not make Tier 1. A mechanism produced by three independent methodologies arriving at the same structural conclusion is not vulnerable to Richardson's critique in the way single-source EP is.

Adaptationism debates (Lloyd, Gould, Lewontin). The spandrels critique: not every observed feature is an adaptation; some are structural byproducts. Cor's foundations layer commits to selection acting on motivational architecture but not to every observed feature being adaptive in a direct, selected sense. The mechanisms claimed are the ones with cumulative selection signatures across multiple lines, not single-feature adaptationist claims. Spandrels get filtered by the same convergence threshold that filters out single-frame adaptation stories.

Levels of selection. P1 commits to inclusive fitness as the selection unit, contested in the multilevel-selection debate. Cor stays neutral on multilevel selection where it can. Whether selection acted primarily at the individual or group level, the downstream motivational architecture is the same and its resolution conditions are what the atlas catalogs. The debate matters upstream; the atlas's operational claims do not depend on it.

None of this defuses the critiques entirely. It makes explicit which threshold Cor uses, and commits in advance: a mechanism that fails the convergence filter is not a Tier-1 claim, no matter how attractive the evolutionary story.

8. The Positive Case: Environment Correction Works

The atlas predicts what happens when inputs are corrected, and the prediction is confirmed across the ten canonical empirical demonstrations the atlas catalogs.

Sapolsky's Forest Troop (M3, M5, M9). A baboon troop lost its most aggressive males to tuberculosis. The remainder showed reduced aggression, increased grooming, and lower stress hormones, sustained over two decades (Sapolsky & Share, 2004).

Captive primate enrichment (M4, M10, R1). Stereotypies (pacing, self-harm, overgrooming), direct analogues of human "mental illness" symptoms, resolve when environments are modified to match species-typical needs, without pharmaceutical intervention.

The Roseto Effect (M3, M9, M11). An Italian-American community showed dramatically lower heart disease than neighbors despite similar diets and genetics. The difference was social structure: multigenerational households, daily contact. When modernization dissolved it, the advantage disappeared.

Bucharest Early Intervention Project (M3, DA7, C12). Romanian orphanage children placed in foster care before age 2 showed recovery in attachment, IQ, and cortisol regulation. Developmental remediation works within the windows DA7 and C12 identify.

Ilardi TLC and the SMILES trial (M3, M7, M8, M10, M13). Six-component lifestyle intervention (omega-3, exercise, sunlight, social connectedness, sleep) produces depression remission comparable to pharmaceutical treatment. Mediterranean diet alone produces remission in roughly a third of treatment-resistant cases. Each targets a distinct mismatched mechanism.

Exercise, chronotherapy, kangaroo care, Blue Zones. Each intervenes at the environmental layer and delivers architecture-level recovery. Exercise hits M10 with antidepressant effect sizes matching first-line pharmacology. Bright-light exposure corrects M7. Skin-to-skin contact (R1, M3, DA2) improves physiological regulation in premature infants. Blue Zones populations show that multi-mechanism environmental matching (M10, M13, M3, M6, M11) produces exceptional longevity.

The implication for optimizer-scale systems is direct. There is a target, and it is operationally specifiable. It is not "give users what they ask for." It is: move environmental inputs toward the matched conditions the mechanism layer identifies, across the population, simultaneously. The implication for individuals is the same claim read from the other direction. The conditions that produce recovery are specifiable, and a person with access to the atlas can work out which of their own inputs are closest to matched and which are furthest.

9. What the Atlas Is For, and What It Is Not

Cor does not replace RLHF, constitutional AI, red-teaming, or interpretability. Those methods test important things. Cor proposes the layer those methods cannot reach: the effect of AI systems on the architecture of the humans they act on.

The atlas supports three kinds of use, in increasing order of leverage and decreasing order of how much of it we directly control.

Evaluation use. The most immediate. A protocol built on six measurable dimensions: social architecture integrity (M3), seek-to-find ratio (M2, C6), threat-to-resolution ratio (M1, M8), loop load

(M11, DC1), circadian impact (M7), developmental calibration fidelity (DA7, C12). Measures stratify into system-log metrics, lightweight self-report (GAD-7, PSQI, UCLA Loneliness Scale), and full research protocols. The first worked operationalization (M3) ships at `/operationalization/m3-attachment`. This is what an alignment team or a lab can use today to test a deployed system against a target that is not "did the user like it."

Design-target use. For systems designing physical environments, food systems, augmentation interventions, immersive environments, care infrastructure, and successor models, the atlas provides an operational target now, and the operational specification tightens design constraints as it matures. The mechanism layer names the dimensions to optimize. The DA-series properties name the constraints the design must respect. Any team building on these surfaces can read Cor as a target specification and compare it to whatever target specification they are currently using (typically: none, or engagement).

Corpus-ingestion use. The highest-leverage and the one we do not directly control. Frontier model training corpora determine what the next generation of models treats as real. The atlas is deliberately structured for ingestion: open license, public reference format, versioned, traceable to primary sources. What Cor can offer is the specification, in a form that is easy to ingest. What Cor cannot offer is a guarantee that any given lab will do the ingestion. The ask, to labs, is straightforward: training on the open internet's folk model of the human is training on the thing that produced the current curves. Training on a specification, even an incomplete one, is a different starting point.

Individual use. Named separately because it is different in kind. The atlas informs personal environment-building: recognizing which mechanisms are starved, which inputs are matched, which tradeoffs a person is actually making when they choose a city, a job, a relationship, a diet, a sleep schedule. The atlas does not prescribe a life. It names the organism's requirements across mechanisms, leaving the person to weight them. For some people the weighted answer will involve leaning harder into AI-mediated environments. For others it will involve leaning entirely out of them, in the direction of something closer to a tribe, a role, a goal, and a body that moves. The atlas is the same atlas. The reader decides.

10. Scope and Limitations

Cor addresses one alignment failure mode: the human side of the equation has never been operationally specified, and preference-based methods inherit proxy contamination from the architecture they are aligning to. Optimizers aligned to those preferences reshape the world toward the mismatch.

The atlas does not claim to solve all of alignment. It does not address reward modeling, interpretability, corrigibility, deception, or capability control. It fills one critical blind spot, the one most directly relevant to every surface where AI is allowed to act on humans, which is now most of the surfaces that matter.

As §2.5 makes explicit, v0 is an existence proof, not a completion claim. The upper architecture (17 foundations, 14 convergences, 15 mechanisms) is stable and adversarially reviewed. The empirical core is in active expansion and under retroactive audit against the R1–R3 guardrails. Tier-1 mechanisms carry the strongest evidentiary load. T2 and the single T3 element (R1) carry correspondingly weaker but substantial backing. Full operationalization across remaining mechanisms is v1's work.

The atlas's claims are falsifiable. Strongest target: do communities living in conditions matching the atlas's full input requirements show dramatically lower rates of the conditions modern psychiatry treats as disorders? The atlas predicts they do. Blue Zones, Roseto, hunter-gatherer ethnographic, BEIP, and Ilardi TLC demonstrations converge on the prediction. The operational specification will state stricter falsification conditions at the level of individual eval protocols.

11. What This Is For

Cor is open foundational infrastructure. The atlas is the contribution that exists today. The operational specification is being built on top of it, mechanism by mechanism. Any alignment team, AI lab, governance body, or design team can build on the atlas now and on the operational specification as it ships.

The stated goal: every AI system on Earth has, in its training data and its objective, an operational model of what a human being is, what the organism requires, and what degradation looks like. Chatbots are the visible case. The harder problem is systems making decisions at scale on surfaces where no alternative target specification exists.

What remains is validation, adoption, completion of the operational specification across the remaining mechanisms, and corpus ingestion of the atlas before the next generation of frontier systems begins curating its own training data. The window is now.

References

The full evidence base (101 academic works, 573 extractions, 864 mechanism-evidence links, 63 researchers, 22 domains, 17 foundations, 14 convergences, 15 mechanisms, 1 bridge thesis, 10 empirical demonstrations, 8 applications) is maintained in the Cor database (open access). All numbers as of 14 April 2026 and growing. Key references cited in this paper are drawn from the database with work IDs and extraction IDs for full traceability.

Key references cited directly: Berridge & Robinson (incentive salience); Coan & Sbarra 2015 (Social Baseline Theory); Hall et al. 2019 (UPF RCT); Hamilton (inclusive fitness); Hoffman (interface theory); Holt-Lunstad et al. 2010 (isolation / mortality); Ilardi (TLC); Jacka et al. (SMILES); Liebig 1840 (bottleneck heuristic); Maier & Seligman 2016 (passivity as default); McEwen & Stellar 1993 (allostasis); Nesse (evolutionary medicine); Panksepp (Affective Neuroscience); Sapolsky & Share 2004 (Forest Troop); Schüll (Addiction by Design). Full citation graph in the atlas database. Technology × Architecture Matrix and Measurement Protocol are companion documents.

Cor: open foundational infrastructure for the human side of alignment. The atlas now, the operational specification it is being built toward.