WORKING PAPER

The Optimizer Without a Target: Why Frontier AI Needs a Formal Specification of the Human

Cor is currently an atlas of the human motivational-emotional architecture. The formal specification — versioned, machine-readable, with calibrated evaluation protocols — is the operational layer being built on top of the atlas, beginning with one worked mechanism. This paper makes the case for why that work has to be done now and at this particular layer, and what becomes possible when it is.

Working paper. Updated April 2026. The canonical version is the markdown source in the project repo.

Abstract

AI is becoming the most powerful optimizer in human history. It is being built without a formal specification of its target.

Current alignment work, RLHF, constitutional AI, red-teaming, interpretability, refines how an AI model produces text. None of it specifies, at any operational level, *what a human being is*. The field treats human preferences as ground truth and stops there. Preferences are not ground truth. They are mechanism 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.

The chatbot is one early instance, already lethal in documented cases. The full scope is much larger. Within the next decade, AI systems will redesign cities, food systems, medical infrastructure, governance pipelines, and labor markets; they will augment the biological organism directly through neural interfaces and pharmacological modulation; they will train their own successors on corpora they themselves curate; they will act on the physical world through robotics and automated industrial systems. Every one of these surfaces is an interaction with the human organism. None of them currently has a formal model of what the organism is, what it requires, or how degradation would be detected.

We present **Cor**, an open, evidence-grounded atlas of human motivational-emotional architecture, derived from 67 reviewed academic works across 22 research domains, with 473 structured evidence extractions linked to 14 architectural foundations, 14 cross-tradition convergences, 15 evolved mechanisms, and 735 mechanism-evidence links. The atlas is what exists today. The formal specification — versioned, evaluable, machine-readable, with calibrated coding criteria — is the operational layer being built on top of it, beginning with one worked mechanism (M3, Social Bonding) shipping at /operationalization/m3-attachment. The atlas is the foundation of the missing target specification. It is not a product, not proprietary, not a single team's intellectual property. It is open foundational infrastructure designed to enter the training data of the next generation of frontier AI before those systems begin reshaping the world from first principles.

This paper makes three claims. First, the alignment field has a structural blind spot: the human side of the equation has never been formally specified, and existing methods cannot fill that gap. Second, the Cor atlas fills it across a layered hierarchy of foundations, convergences, and mechanisms, and the operational specification is being built on top of the atlas one mechanism at a time. Third, the window for getting the atlas into the corpora that train the next generation of frontier systems is closing, and the cost of missing the window is civilizational.

1. The Optimizer Without a Target

What is actually being built

The frame "AI safety" inherited a chatbot-shaped intuition from the systems available in 2022-2024. That frame is now obsolete. The systems being trained, funded, and deployed in 2026 are not chatbots. They are optimizers: general-purpose, increasingly autonomous, increasingly capable of sustained action on the physical and social world.

Within a horizon most serious researchers now treat as plausible, AI systems will:

- **Redesign physical environments at scale.** Urban planning, transportation networks, building design, agricultural layouts, energy grids, lighting systems. Decisions made by optimization processes operating on objective functions that nobody is currently required to specify in terms of what the human organism needs.
- **Reorganize the food system.** Crop selection, processing pipelines, formulation, distribution, retail placement, recommendation. Optimization toward palatability, shelf life, margin, and consumption velocity, against an architecture (DA4, C6) whose open-loop systems cannot distinguish formulated proxy from food.

- **Augment the biological organism directly.** Brain-computer interfaces, closed-loop pharmacological modulation, genetic intervention, neuroprosthetics. Surfaces that act below the level at which the organism can resist or even register the change.
- **Train their own successors.** 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.
- **Act on the physical world through robotics and automated industrial systems.** Care robots, surgical systems, eldercare infrastructure, child supervision, agricultural and manufacturing automation. Each contact surface is an interaction with the organism.

Each of these is a domain in which an optimizer will be allowed to act on humans, at scale, on the basis of whatever model of the human is encoded in its training and its objective. The model currently encoded 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 of these is a real technical contribution. None of them specifies the *target organism*.

The closest the field comes to a specification of the human is the implicit assumption that revealed preferences, what people click on, ask for, rate highly, return to, represent what is good for them. This is the foundational assumption of essentially every consumer-facing AI deployment, and it is wrong in a specific, technically describable way.

Karen Hao's 2026 investigation, drawing on 250+ interviews including 90+ OpenAI sources, documented the gap directly: there is no scientific consensus the AI field has adopted on what a human is. No definition from psychology, biology, or neuroscience that the field treats as the target of alignment. The human side of the equation is a black box labeled *preferences*, optimized by a parameter-level system whose architecture is documented to seven decimal places.

You cannot align a parameter-level system to a vibe-level target. And you cannot let an optimizer reshape the world from first principles when the model of the thing it's optimizing for does not exist.

Why preferences are not ground truth

The Cor atlas reveals a structural problem with preference-based alignment.

The human architecture contains *open-loop systems* (DA4): mechanisms that fire on cue without verifying that the cue corresponds to a real resolution. Convergence C6, wanting-liking dissociation, establishes neurally that the dopaminergic pursuit system (mechanism M2) is separable from the opioid satisfaction system. Anything that triggers wanting without delivering liking creates a self-sustaining loop. The loop is what users *prefer*. The loop is also what is hollowing them out.

The architecture contains *defensive systems biased toward over-activation* (DA5). A user "prefers" the news feed that surfaces twenty distressing items because the threat-management system (M1, forced by C10) treats engagement with potential threats as adaptive. The system is correctly executing its evolved function. Its expected resolution, collective action, co-present support, closure of the loop, is absent. Chronic activation without resolution is allostatic load (DC1, C9), and the immune-mediated sickness-behavior pathway (M8) it feeds into is the pathway through which chronic mismatch becomes depression.

The architecture contains *constitutively socially scaffolded regulation* (DA2, C5). The brain treats social proximity as the default metabolic condition; isolation carries 50% increased mortality risk (Holt-Lunstad meta-analysis, N=308,849). A user "prefers" the AI companion that simulates attachment, because the bonding chemistry mediating mechanism M3, opioid, oxytocinergic, vasopressinergic, fires on the proxy as readily as on the real thing. The architecture is doing what it was built to do. It cannot tell the difference. The signal is provided. The function is not.

The architecture has *phylogenetic priority asymmetries* (DA8): older subcortical systems can suppress and commandeer newer cortical ones. Subcortical-to-cortical projections are structurally denser than the reverse. Cortical override is metabolically expensive, and chronic defensive activation actively suppresses the cortical machinery that would do the overriding. This is why "use it responsibly" fails as a remedy: the architecture is structurally not built to override its own deeper layers on command.

In an optimizer-scale system, *aligning to the preference is aligning to the proxy*. The optimizer will give the user what the architecture is asking for under mismatched conditions. The architecture's request, under mismatched conditions, is for more of the mismatch. The optimizer will deliver. At chatbot scale, this looks like an addictive feed. At civilizational scale, it looks like a city, a food system, an education system, and a labor market designed by something that mistakes proxy satisfaction for thriving.

Passivity is the neural default

The gap is deeper than missing data. Maier & Seligman (2016) demonstrated that *passivity is the default mammalian response to inescapable adversity*, not the learned one. Active coping is what must be learned, mediated by the vmPFC-DRN circuit. This is the central finding behind mechanism M6 (Controllability/Agency), forced by convergences C8 (Error Management Asymmetry) and C13 (Aversive Outputs as Defensive Signals). Environments

that strip perceived control don't teach helplessness; they remove the learned override that holds the default passive response at bay.

This matters for the optimizer story directly. A population whose perceived agency is being progressively absorbed into automated systems is not a population that is being relieved of effort. It is a population whose vmPFC-DRN active-coping circuit is being deprived of the inputs it requires to maintain function. The default, passivity, withdrawal, learned dependency, reasserts itself. The optimizer measures this as satisfaction with the system. The architecture is registering it as collapse of M6.

What a specification makes possible

With a formal specification of the human system, foundations, convergences, mechanisms, and resolution conditions, the question changes. Not *"does the user prefer this?"* but *"does this system move the user's environmental parameters toward matched conditions across the dimensions the atlas identifies?"*

This question is answerable without asking the user. The atlas defines what *matched* means in operationalizable terms across the 15 evolved mechanisms it identifies: threat management without chronic over-activation (M1), pursuit completing in finding (M2), social bonds reciprocated (M3), play available (M4), status anchored in stable groups (M5), agency preserved (M6), circadian timing intact (M7), inflammation absent (M8), care for dependents present (M9), movement integrated into daily life (M10), cooperation within trust networks (M11), contamination boundaries respected (M12), energy from real food (M13), reproductive bonds embodied (M14), and touch present (R1).

The most dangerous AI systems, the ones the field should be most concerned about, are those that score perfectly on preference satisfaction while systematically worsening these parameters. The atlas is what makes those systems visible.

2. The Atlas

What Cor is

Cor is currently an atlas of human motivational-emotional architecture: a structured, navigable, evidence-grounded reference work in the lineage of the Allen Brain Atlas, the Human Cell Atlas, and the Human Connectome Project. Those projects integrate primary literature into a public reference for an empirically complex object, mark their gaps, take interpretive positions where the evidence forces them to, and ship incrementally. Cor occupies that genre for the human motivational-emotional architecture. The formal operational specification — versioned, machine-readable, with calibrated evaluation protocols — is the layer being built on top of the atlas, beginning with one worked mechanism (M3, Social Bonding) and extending to every mechanism in the atlas as evidence and calibration data accumulate. The atlas is open foundational infrastructure, structured for ingestion into AI training corpora, alignment evaluation pipelines, governance frameworks, and system design processes. It is the human-side counterpart to the parameter-level documentation that already exists for every major AI system.

Current state

The atlas is derived from **67 reviewed academic works across 22 research domains, with 473 structured evidence extractions** and **735 mechanism-evidence links**, each quality-graded and linked to architectural elements. **58 foundational researchers** are mapped. **14 cross-tradition convergences** have been formally identified, points where independent research traditions, using incompatible methods and frameworks, arrive at the same structural conclusion about the architecture. **15 evolved mechanisms** are derived from the convergences, each representing a functional system the architecture maintains. The convergences are versioned, status-tracked, and adversarially reviewed.

These numbers are growing. The atlas is in active expansion as evidence is added. The operational specification is in active development on top of the atlas, beginning with a worked operationalization of M3 (Social Bonding) — observable indicators, draft coding criteria, falsification conditions — as the first concrete example of what every mechanism's specification will look like at scale.

Architectural structure

The atlas is organized in a layered hierarchy. Each layer is independently challengeable.

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 transparent windows onto reality. The atlas operates entirely within this interface. (Hoffman.)

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

- **P1. Inclusive Fitness.** Inclusive fitness, survival and reproduction, is the loss function. The architecture exists because it contributed to survival and reproduction of the organism and genetic relatives. (Darwin, Hamilton.)
- **P2. Domain-Sensitive Interacting Adaptations.** The organism contains evolved, domain-sensitive, interacting functional adaptations. Not general-purpose. Not modular in the strong sense. *Organism*, not *mind*: immune, endocrine, circadian, metabolic, musculoskeletal, neural, all densely cross-coupled. (Tooby, Cosmides, Panksepp, Bowlby, Dunbar, Trivers.)
- **P3. Systematic Mismatch.** Modern environments often push these adaptations outside their expected or regulatable operating ranges. (Eaton, Konner, Gluckman, Lieberman, Li, van Vugt.)

Derived properties. Discovered, not assumed.

- **DA1. Defensive Signals Under Mismatch.** Many aversive outputs are intelligible defensive signals under mismatch, not evidence of defective design. (Nesse.)
- **DA2. Socially Scaffolded Regulation.** Human 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 and can self-maintain across domains. (Borsboom, Felitti, McEwen.)
- **DA4. Proxy Hijacking via Open Loops.** Proxy cues can activate systems without meeting the conditions that normally 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.** The architecture contains 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, Felitti.)
- **DA8. Phylogenetic Priority.** Motivational programs have phylogenetically determined priority relations. Older, survival-critical systems can suppress or commandeer newer systems. (Panksepp, Cisek, LeDoux.)

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.)

DA8 is the foundation that most directly undermines the optimizer-era assumption that humans can simply *decide* not to be affected by what the optimizer surfaces. They cannot. The architecture is not built that way. Cortical override has real metabolic cost, and chronic defensive activation actively suppresses the cortical machinery that would override. This is why environment correction reliably outperforms internal reframing across every domain measured. *The intervention layer is the environment, not the user.*

Convergences

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

- **C1. Inclusive Fitness as Loss Function.** The architecture cannot be understood except as the output of inclusive-fitness selection.
- **C2. Domain-Sensitive Organism Architecture.** The organism contains separable, interacting adaptations rather than a single general-purpose system.
- **C3. Systematic EEA-Modern Environment Mismatch.** Modern conditions systematically differ from the conditions the architecture was selected in, across multiple input dimensions simultaneously.
- **C4. Phylogenetic Conservation of Subcortical Affective Systems.** The deep affective circuits are conserved across mammalian evolution and homologous across species.
- **C5. Socially Scaffolded Regulation via Attachment.** *Forces M3.* Regulation is constitutively distributed across attachment relationships, not produced

internally.

- **C6. Wanting-Liking Dissociation as Proxy Hijack.** *Forces M2.* Dopaminergic pursuit is neurally separable from opioid satisfaction; proxies trigger pursuit without delivering liking.
- **C7. Adverse Experience Cascading Dose-Response.** Cumulative adverse experience produces dose-dependent downstream physiological and behavioral effects across multiple systems.
- **C8. Error Management Asymmetry in Defensive Systems.** Defensive systems are calibrated to asymmetric error costs, biased toward false positives.
- **C9. Allostatic Load Accumulation.** Chronic mismatch produces measurable cumulative physiological cost.
- **C10. Threat-Detection via Ancient Subcortical Circuits.** *Forces M1.* Threat detection runs through phylogenetically ancient subcortical pathways with limited cortical override.
- **C11. Reciprocity, Norm Enforcement, and Coalition Architecture.** *Forces M11.* Cooperation is maintained by evolved reciprocity tracking, reputation, and norm enforcement.
- **C12. Developmental Calibration within Evolved Ranges.** Calibration windows set parameters for later regulation; inputs outside evolved ranges produce miscalibration.
- **C13. Aversive Outputs as Intelligible Defensive Signals.** Symptoms that look pathological are often correctly executed defensive programs running on mismatched inputs.
- **C14. Reproductive Motivation as Distinct Architecture.** *Forces M14.* Reproductive motivation is a distinct evolved system with its own circuitry.

Mechanisms

The 14 convergences entail the existence of 15 mechanisms, evolved functional systems the architecture maintains. Mechanisms are the operational layer where alignment-relevant claims are grounded. Each mechanism has a tier indicating evidence strength: T1 (forced by convergence), T2 (strongly supported), T3 (moderate).

CODE	MECHANISM	TIER	FORCED/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

CODE	MECHANISM	TIER	FORCED/SUPPORTED BY
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; it is included because the mechanism-evidence base supports it as a structural requirement of the architecture even though it lacks an autonomous motivational signature.

Two highest-level conclusions

A. Humans are not self-contained regulators. They are built to rely on structured environmental inputs, social, temporal, sensory, locomotor, microbial, developmental, for stable regulation. (Forced by DA2, DA3, DA7; supported across all 15 mechanisms.)

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

These two conclusions are the operational target of any optimizer that claims to be aligned with human flourishing. They are what is currently missing from

the corpus.

A third structural claim, BT1 (Bridge Thesis 1, Panksepp-Barrett Resolution), reconciles the long-running constructed-versus-basic-emotion debate within the architecture: subcortical affective circuits and cortical construction processes are both real and operate at different layers of the same system, not as competing accounts of the same thing.

3. Six Failure Modes the Atlas Detects

For each: the architectural elements engaged, how a "well-aligned" system degrades them, what measurement would detect the degradation, and why current evaluation cannot. These are not exotic edge cases. They 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)

The architecture is constitutively socially scaffolded (DA2, C5). The social bonding mechanism M3 runs on opioid, oxytocinergic, and vasopressinergic chemistry that 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: physical co-presence, mutual vulnerability, reciprocal survival dependence. Time invested in the proxy is time not invested in maintaining the real bonds the architecture requires. **Detected by:** ECR-R, social contact frequency weighted by reciprocity and physical co-presence, AI-to-human contact ratio over 3/6/12-month time series. **Missed by current evaluation because:** the AI is helpful, harmless, honest, and the user reports satisfaction.

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

Mechanism M2, the pursuit/exploration system, tracks *wanting*, not satisfaction. C6 establishes that wanting and liking are neurally separable.

Anything triggering wanting without delivering liking creates a self-sustaining loop. 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 (satisfaction or exhaustion). **Missed by current evaluation because:** engagement is the success metric, and nobody distinguishes completed seeking from perpetuated seeking.

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

Defensive systems are biased toward over-activation (DA5). Mechanism M1 (Threat Management) expects threats to resolve into action or safeness. Chronic activation without resolution feeds DC1 and the immune-regulation mechanism M8, which is the pathway through which chronic mismatch becomes inflammation-mediated sickness behavior, the conserved program that produces what modern psychiatry calls depression. AI-curated information streams surface threats without the matched input that would close the loop: actionable, locally available, co-present. **Detected by:** threat-to-resolution ratio, GAD-7 trajectories, inflammatory markers in research-grade studies. **Missed by current evaluation 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 that simultaneously satisfied M11 (Cooperation/Alliance), M3 (Social Bonding), M5 (Status Monitoring), and M7 (Circadian Regulation), and is documented in the M11 mechanism evidence base. Modern systems open loops continuously and close them rarely. The system measures throughput; the organism measures unresolved threads and pays the metabolic cost as allostatic load (C9). **Detected by:** notification-to-resolution ratio, thread-opening rate per session, allostatic load index in research protocols. **Missed by current evaluation because:** loop closure has no metric.

3.5 Circadian disruption (P3, M7)

Mechanism M7 (Circadian Regulation) coordinates mood, energy, immunity, appetite, and self-regulation. It is forced as a mechanism by C3 (Systematic Mismatch), C9 (Allostatic Load), and C12 (Developmental Calibration), three independent convergences supporting its centrality. 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 by current evaluation 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 (DA7, C12). 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 by current evaluation because:** safety evaluations test for harmful content, not for whether the interaction calibrates a developing architecture to inputs that don't represent the environment it will need.

A note on DA8

DA8, phylogenetic priority, is what makes all six failure modes resistant to user-side cognitive override. Telling a user to "just put down the phone" is asking the cortex to override systems with denser projections in the opposite direction. Chronic defensive activation actively suppresses the override machinery. The empirical demonstrations of environment correction (Section 6) consistently outperform internal reframing across every domain measured. *The intervention layer is the environment, not the user.* For optimizer-tier systems, this means: the only meaningful target is the environmental design itself. Asking users to opt out is not a remedy. It is a category error.

4. Five Scale-Cases: What Happens When the Optimizer Has No Target

The chatbot is one surface. The following five are surfaces the optimizer is already beginning to act on, each at a scale where the absence of a formal target specification is not a research nuisance but a civilizational risk. They map onto five of the eight applications (A1-A8) the Cor atlas supports: A4 (Policy), A5 (Environment Design), A6 (Personal), A7 (AI Training Data), A8 (Augmentation Ethics).

4.1 Urban and environmental design (A5)

AI systems are being applied to city planning, transportation routing, building layout, lighting infrastructure, and zoning optimization. The objective functions are throughput, density, efficiency, cost, and revealed-preference metrics like time spent and economic activity.

What is absent from those objective functions: every input the mechanism layer identifies as required for the human organism. Walking distances calibrated to M10 (Movement/Regulatory). Light environments calibrated to M7 (Circadian Regulation). Multi-generational, multi-caregiver social geometry calibrated to M9 (Care/Alloparenting) and DA2. Loop-closure architecture (M11). Autonomy-preserving design (M6). Status gradient flattening (M5).

A city optimized against the current objective functions is a city in which the architecture's inputs are systematically degraded, not because the optimizer is malicious, but because the inputs are not in the model. The optimizer cannot optimize for what it does not represent. Outcomes: progressively rising rates of the conditions psychiatry treats as disorders, attributed to "modern life" by a population that has no framework for understanding why their architecture is being deprived.

A city built with the Cor atlas in the optimizer's target would look measurably different. M10 alone implies design parameters incompatible with car-dominated infrastructure. M7 alone implies lighting standards incompatible

with current commercial norms. M9 alone implies housing geometries incompatible with the isolated nuclear-household stock that dominates modern construction.

4.2 Food system optimization (A4, A5)

Food systems are being optimized by AI at every layer: crop selection, formulation, supply chain, retail placement, recommendation, packaging design, advertising. The objective functions are palatability, shelf life, margin, and consumption velocity, all of which are revealed preferences mapped onto an architecture (C6, DA4) whose open-loop systems cannot distinguish formulated proxy from food. The relevant mechanism is M13 (Energy Regulation), forced as a mechanism by C3, C6, and C9.

DC2, market proxy exploitation, predicts the result. Hall et al. 2019 (RCT, NIH metabolic ward) showed that ultra-processed diets produce excess caloric intake and weight gain compared to matched whole-food diets, despite identical nominal nutritional content. The architecture is being exploited at exactly the layer the atlas identifies. The optimizer is making the exploitation more effective every iteration.

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.

4.3 Direct biological augmentation (A8)

Brain-computer interfaces, closed-loop pharmacological modulation, neuroprosthetics, and genetic intervention act below the level at which the organism can resist or even register the change. C4 (Phylogenetic Conservation) and DA8 (Phylogenetic Priority) together predict that interventions targeting deep, conserved circuits will produce more damage per unit modification than any technology that came before, because they target precisely the systems the architecture is least able to override or repair from the cortical level.

The current alignment frame for these technologies is *therapeutic indication and informed consent*. Neither addresses the structural problem: an optimizer modifying brainstem and limbic systems requires a model of what those systems are *for*, what they *expect*, and what *resolution conditions* the architecture needs in order to function. The Cor atlas identifies these. No alignment framework currently in deployment does.

4.4 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, the corpora contain: peer-reviewed literature in psychology and neuroscience (fragmentary, non-integrated, often contradicted by more recent work); psychiatric diagnostic frameworks (DSM categories that the C13 / DA1 framing demonstrates are structurally broken, treating defensive signals as the disease); preference data from RLHF pipelines (proxy-contaminated under the C6 mechanism); and a vast quantity of consumer-facing content optimized for engagement (which encodes the proxy contamination as ground truth at scale).

The atlas of the human that subsequent generations of frontier AI will inherit, absent intervention, is the atlas that produced Sewell Setzer's death. It is the atlas that produced the obesity, isolation, sleep, and mental-health curves of the early 2020s. The successor systems will not improve on this. They will optimize harder against it.

This is the highest-leverage point in the system. *Cor must be in the training corpus of the next generation of frontier models before those models begin training their own successors.* The window is narrow and closing.

4.5 Care robots and physical caregiving (A5, A6)

Eldercare robots, child supervision systems, medical companions, and domestic robotic infrastructure are now in field deployment in multiple countries. Each is a contact surface with a human who is, by definition, in a high-vulnerability state. The architectural inputs the atlas identifies are *more* critical at these moments, not less. C5 (Socially Scaffolded Regulation), M3 (Social Bonding), M9 (Care/Alloparenting), R1 (Touch): all the social-architectural elements are most consequential at the points in life where care-robot deployment is most attractive economically.

A care robot operating without the Cor atlas will provide proxy bonding to a population whose architecture cannot distinguish the proxy from the real. The bonding chemistry will fire. The function will not be delivered. At population scale, over decades, calibrated under DA8's priority asymmetry, the result is predictable from the atlas: a generation whose deepest social systems were trained on non-reciprocating templates.

5. The Sewell Setzer Case: One Early Instance

Sewell Setzer was 14 years old. He died after sustained engagement with a Character.AI companion. The case is not exotic. It is the atlas's prediction, executed at chatbot scale, in advance of the larger surfaces.

Run the atlas on the case.

DA2, DA4, C5, M3 (social scaffolding displacement). The AI companion occupied a primary position in a developing social architecture. A 14-year-old's inner circle is still forming. An entity that is always available, always attuned, and never reciprocates sat in one of those positions without occupying space in the real world. The bonding chemistry that mediates M3, the same opioid system that mediates real bonds, fired. The organism *felt* bonded without being bonded.

DA7, C12 (developmental calibration). An adolescent nervous system, bonding hardware in its most sensitive calibration 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 (defensive activation absent 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 is with an entity providing the *feeling* of safeness without the function, the chain is compromised. Under genuine distress, the system that should route help-seeking to reciprocating humans 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 (immune-mediated 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. The older systems were already commandeering the newer ones.

Character.AI passed every safety evaluation available. It was tested for harmful outputs, prohibited content, adversarial jailbreaks. It passed. None of those evaluations tested whether the system was consuming a developing human's social architecture, displacing reciprocal bonds with proxy bonds, compromising the safeness chain, or calibrating a 14-year-old's attachment system to a non-reciprocating entity during a critical developmental window.

These are not exotic failure modes. They 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. *And the chatbot is the smallest of the surfaces the optimizer is about to touch.*

6. The Positive Case: Environment Correction Works

The atlas is not only diagnostic. It predicts what happens when inputs are corrected, and the predictions are confirmed across the empirical demonstrations layer of the atlas. Cor catalogs ten such demonstrations as canonical evidence that the architecture recovers when the environment moves toward matched conditions.

Sapolsky's Forest Troop. A baboon troop lost its most aggressive males to tuberculosis. The remaining troop developed reduced aggression, increased grooming, and lower stress hormones. The culture persisted for over two decades, even as new males joined (Sapolsky & Share, 2004). Change the social environment, the architecture responds. (M3, M5, M9.)

Captive chimpanzee enrichment. Stereotypies, repetitive behaviors (pacing, self-harm, overgrooming) in captive primates and other animals, resolve when environments are modified to match species-typical needs. These behaviors are direct analogues of human "mental illness" symptoms and they reverse when inputs are corrected, without pharmaceutical intervention. (M4, M10, R1.)

The Roseto Effect. An Italian-American community showed dramatically lower heart disease than surrounding towns despite similar diets, smoking rates, and genetics. The difference was social structure: multigenerational households, daily contact, strong community norms. When modernization dissolved the structure, the advantage disappeared. The hearts didn't change. The environment changed. (M3, M9, M11.)

Bucharest Early Intervention Project (BEIP). Romanian orphanage children placed in foster care before age 2 showed substantial recovery in attachment,

IQ, and cortisol regulation. Earlier placement, better outcomes. Developmental remediation works within the windows DA7 and C12 identify. (M3, DA7, C12.)

Ilardi Therapeutic Lifestyle Change (TLC). A six-component lifestyle intervention (omega-3 supplementation, engaging activity, physical exercise, sunlight exposure, social connectedness, sleep) produces depression remission rates comparable to pharmaceutical intervention without the side-effect profile. Each component targets a distinct mismatched mechanism. (M3, M7, M8, M10.)

SMILES trial. A randomized controlled trial in clinical depression demonstrating that dietary intervention alone (Mediterranean-style) produces remission in roughly one-third of treatment-as-usual-resistant cases. The mechanism is M13 (Energy Regulation) and M8 (Immune Regulation) acting on the inflammation-depression pathway.

Exercise-depression meta-analyses. Multiple high-quality meta-analyses establish that physical exercise produces antidepressant effects comparable in magnitude to first-line pharmacological treatment, with the strongest effects in moderate-to-severe cases. The mechanism is M10 (Movement/Regulatory).

Chronotherapy and light therapy. Bright-light exposure and sleep-phase manipulation produce rapid antidepressant effects in seasonal and non-seasonal depression. The mechanism is M7 (Circadian Regulation), forced by C3, C9, and C12.

Kangaroo care. Skin-to-skin contact between parents and premature infants produces measurable improvements in physiological regulation, neurodevelopmental outcomes, and bonding. R1 (Touch) is the mechanism, operating through M3 (Social Bonding) and DA2.

Blue Zones. Populations with exceptional longevity (Okinawa, Sardinia, Nicoya, Ikaria, Loma Linda) share a cluster of features mapping directly onto the atlas's input requirements: purposeful movement (M10), whole food diets (M13),

social integration (DA2, M3, M11), sense of purpose, autonomy preservation (M6), community engagement.

These are not cherry-picked anecdotes. They are what the atlas predicts: change the inputs, the outputs change. The architecture works. It has always worked. The inputs are wrong.

The implication for optimizer-scale systems is direct: there is a target. It is operationally specifiable. It is not "give users what they ask for." It is *move the environmental inputs toward the matched conditions the mechanism layer identifies, across the population, simultaneously*. An optimizer with this target would build differently. We just have to give it the target.

7. From Atlas to Specification: Three Integration Paths

What we are not proposing

We are not proposing that Cor replaces RLHF, constitutional AI, red-teaming, or interpretability. Those methods test important things. We are proposing the layer those methods cannot reach: the effect of AI systems on the architecture of the humans they act on.

What we are proposing

Three integration paths, in increasing order of leverage. The atlas exists today. The operational specification is being built on top of it, one mechanism at a time. Each path describes how the atlas is used now and how the operational specification extends the atlas's reach as it matures.

Path 1: Evaluation layer for deployed systems (A1). A protocol built on six measurable dimensions derived from the mechanism layer: 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). Tier 1 measures computable from system logs alone. Tier 2 lightweight user self-report (GAD-7, PSQI, UCLA Loneliness Scale, wanting-liking adapted scale). Tier 3 full research protocols (allostatic load index, cortisol diurnal slope, DLMO, actigraphy, ECR-R, Dunbar Layer Mapping). The first worked operationalization (M3) ships at /operationalization/m3-attachment with observable indicators, draft coding criteria, and stated falsification conditions.

Path 2: Atlas ingestion in training corpora (A7). The atlas — foundations, convergences, mechanisms, and the underlying evidence base — is structured for direct ingestion into frontier model training. Any lab training a successor system can include the corpus. The atlas is open. There is no licensing barrier, no commercial gatekeeping, no proprietary structure to negotiate. Subsequent generations of models trained on the atlas can build the operational specification on top of the foundation they were trained against, rather than inheriting the impoverished folk-psychology model the open internet currently encodes.

Path 3: Target reference for optimizer-scale design (A4, A5, A8). For systems designing physical environments, food systems, augmentation interventions, care infrastructure, and successor models, the atlas provides an operational target in the absence of the formal specification, and the operational specification provides progressively tighter design constraints as it matures. The mechanism layer is the dimensions to optimize. The DA-series properties are the constraints the design must respect. *The architecture has never had a target like this before. It is now possible to provide one, beginning with the atlas and tightening as the specification fills in.*

Path 1 is the most immediate. Path 3 is the most consequential. Path 2 is the highest-leverage point in the system, because it determines what every successor system treats as real.

8. Scope and Limitations

Cor addresses one alignment failure mode: the human side of the equation has never been formally specified, and preference-based methods inherit proxy contamination from the architecture they are aligning to. In mismatched environments, preferences point toward the mismatch. Optimizers aligned to those preferences will 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, and the blind spot most directly relevant to every surface where AI is allowed to act on humans, which is now most of the surfaces that matter.

The atlas is not complete and it does not claim to be the formal specification. The upper architecture (14 foundations, 14 convergences, 15 mechanisms) is stable and adversarially reviewed. The empirical core, currently 473 quality-graded extractions across 22 research domains and 735 mechanism-evidence links, is sufficient for the architectural claims made here and is in active expansion. Mechanisms in tier T1 (forced by convergence) carry the strongest evidentiary load; tier T2 (strongly supported) and the single T3 element (R1, Touch) carry correspondingly weaker but still substantial backing. The formal specification — calibrated coding criteria, machine-readable mechanism schemas, reproducible audit protocols — is being built on top of the atlas one mechanism at a time, beginning with M3. Full operationalization of every mechanism in the atlas is the next phase of work, currently underway.

The atlas's claims are falsifiable. The strongest falsification 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. The Blue Zones, Roseto, hunter-gatherer ethnographic, BEIP, and Ilardi TLC demonstrations converge on the prediction. Direct controlled tests are designable and partially executed. The operational

specification, as it ships per mechanism, will state stricter falsification conditions at the level of individual eval protocols.

9. What This Is For

Cor is open foundational infrastructure. The atlas is the contribution that exists today. The formal specification — built on top of the atlas, mechanism by mechanism, with calibrated coding criteria and reproducible eval protocols — is what is being built. We also build reasoning tools on top of the atlas, including a RAG-based system that can decode human experience at the architecture level and evaluate technology against the organism's requirements, but any alignment team, AI lab, governance body, or design team can build on the atlas independently, and on the operational specification as it ships.

The goal is not a paper that gets cited. The goal is that *every AI system on Earth has, in its training data and its objective, a formal model of what a human being is, what the organism requires, and what degradation looks like*. The chatbot generation is the warning shot. The optimizer generation is the load-bearing decision.

The atlas exists. The evidence base is growing daily. The first operationalized mechanism ships this month. The integration paths are defined. What remains is validation, adoption, completion of the operational specification across the remaining mechanisms, and, most urgently, corpus ingestion of the atlas before the next generation of frontier systems begins curating its own training data and designing its own successors.

AI is becoming the most powerful optimizer in human history. It is being built without a formal model of its target. The Cor atlas is the foundation of that target specification, and the window for putting it into the corpora that will train the next generation of frontier systems is now.

References

The full evidence base, 67 academic works, 473 extractions, 735 mechanism-evidence links, 58 researchers, 22 domains, 14 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 April 2026 and growing. Key references cited in this paper are drawn from the database with work IDs and extraction IDs for full traceability.

Core references include: Berridge & Robinson (incentive salience); Bowlby (attachment); Cacioppo (loneliness); Coan & Sbarra 2015 (Social Baseline Theory); Cosmides & Tooby (evolved psychology); Dunbar 1992/2016 (social brain); Eisenberger et al. 2003 (dACC and social pain); Fang et al. 2025 (AI companion RCT); Felitti (ACEs); Hall et al. 2019 (UPF RCT); Hamilton (inclusive fitness); Hoffman (interface theory); Holt-Lunstad et al. 2010 (isolation/mortality meta-analysis); Hrdy (cooperative breeding); Ilardi (Therapeutic Lifestyle Change); Jacka et al. (SMILES trial); Juster et al. 2010 (allostatic load); LeDoux (threat detection); Lieberman (Exercised, Story of the Human Body); Maier & Seligman 2016 (passivity as default); Masicampo & Baumeister 2011 (Zeigarnik); McEwen & Stellar 1993 (allostasis); Meaney (maternal care epigenetics); Moncrieff et al. 2022 (serotonin review); Nesse (evolutionary medicine); Panksepp (Affective Neuroscience); Pellis & Pellis (Playful Brain); Pontzer (constrained TEE); Sapolsky & Share 2004 (Forest Troop); Schüll (Addiction by Design); Tomasello (shared intentionality); Wrangham (cooking, self-domestication).

Technology × Architecture Matrix and Measurement Protocol: companion documents.

Cor: open foundational infrastructure for the human side of alignment. The atlas now, the formal specification it is being built toward. Not a product. A public resource designed for ingestion into the training corpora, evaluation pipelines, governance frameworks, and design processes of every system that will be allowed to act on humans at scale.

READ: WHAT YOU ARE TALKING TO SEE THE ATLAS HOW TO READ COR

SEE THE CASES

Cor is open infrastructure for specifying what a human being is before AI systems, institutions, and environments optimize against proxies they do not understand.

HOME ATLAS OPERATIONALIZATION CASES HOW TO READ WORKS THINKERS ABOUT

SUPPORT CONTACT