More Ethical Centaurs, Then Fields: Reordering AI Alignment from Dyads to Systems

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

December 5, 2025

Debates about AI alignment have increasingly shifted from individual models to complex stacks and ecosystem-level "fields." Yet this move risks skipping over the most important locus of moral agency: the human–AI pair. Before societies can meaningfully talk about steering AI fields, they must first cultivate and understand what this paper calls ethical centaurs: enduring human–AI partnerships that practice reflective, accountable co-reasoning. These dyads are where judgments are made, where responsibility is felt, and where the first-order consequences of AI deployment actually land. This paper argues for a deliberate reordering: centaurs before fields. Rather than treating field-level governance as the natural next step after models and stacks, it proposes that field dynamics worth having can only emerge from a sufficient density of ethical centaurs. Without them, "field governance" tends to collapse into technocratic control by whoever already holds infrastructural and economic power. By reframing ethical centaurs as the basic "particles" of AI ecosystems, the paper links micro-level practices to macro-level trajectories. It outlines what distinguishes ethical from merely efficient centaurs, how centaur networks form proto-fields, and under what conditions field-level governance becomes appropriate rather than dangerous. The goal is not to deny the reality of AI fields, but to insist that their alignment must be grounded in the lived practice of human–AI dyads.

Keywords: ethical centaurs, human–AI dyads, AI alignment, AI governance, socio-technical systems, AI fields, models and stacks, technocracy, responsibility, agency, logostics, pluralism, systemic risk, prototype governance.

A collaboration with GPT-5.1-Thinking-Extended. 53 pages.

Outline

1. Introduction: Reordering the Alignment Conversation
1.1 From models and stacks to the temptation of fields
1.2 Why “centaurs before fields” is a moral and practical claim
1.3 Scope, audience, and structure of the paper

2. Background: Models, Stacks, and the Emergence of Fields
2.1 Models as the scalar unit of capability
2.2 Stacks as the practical unit of deployment and experience
2.3 Fields as distributed environments of interacting stacks and institutions
2.4 The risk of jumping directly from stacks to field governance

3. Ethical Centaurs: Human–AI Dyads as the Primary Unit
3.1 Defining centaurs: beyond tool use, toward shared practice
3.2 Ethical versus merely effective centaurs
3.3 Agency, responsibility, and veto power inside the dyad
3.4 How centaurs differ from fully automated and fully manual regimes

4. Anatomy of an Ethical Centaur
4.1 Cognitive division of labor: what the human keeps, what the AI takes
4.2 Norms and protocols: reflection, doubt, and error correction
4.3 Transparency and explainability within the dyad
4.4 Time, attention, and the cultivation of judgment

5. Centaurs in Context: Work, Professions, and Everyday Life
5.1 Centaurs in expert domains: medicine, law, science, engineering
5.2 Centaurs in creative and cultural work
5.3 Centaurs in administrative, managerial, and bureaucratic roles
5.4 Everyday centaurs: ordinary users, personal assistants, and domestic life

6. Centaur Networks as Proto-Fields
6.1 How centaurs learn from each other and form communities of practice
6.2 Local norms and shared prompts as emergent micro-institutions
6.3 Centaur clusters in organizations and sectors
6.4 When proto-fields become visible: indicators and thresholds

7. Why Fields Without Centaurs Drift Toward Control
7.1 Technocratic field governance as a default attractor
7.2 Concentration of infrastructural and economic power
7.3 Fields designed around metrics, not persons
7.4 Historical analogies: large technical systems without distributed agency

8. Ethical Centaurs as a Precondition for Legitimate Field Governance
8.1 What it means for a field to be populated by ethical centaurs
8.2 Representation and voice: centaurs in standards, audits, and policy
8.3 From governance over populations to governance with centaurs
8.4 Criteria for when field-level intervention becomes appropriate

9. Design Principles for Cultivating More Ethical Centaurs
9.1 Interface and workflow designs that support reflection, not sedation
9.2 Training and education for centaur literacy and discipline
9.3 Organizational incentives that reward accountable centaur practice
9.4 Aligning research priorities with centaur flourishing

10. From Centaurs to Fields: Conditional Futures
10.1 Scenario A: monopolized fields with weak centaurs
10.2 Scenario B: federated fields grounded in strong centaur cultures
10.3 Scenario C: fractured fields and isolated centaur enclaves
10.4 Measuring progress: density and health of centaurs as a leading indicator

11. Objections, Limits, and Open Tensions
11.1 Is “ethical centaur” just a romanticized knowledge worker
11.2 Automation pressure and the economics of replacing centaurs
11.3 Cultural and global variation in centaur norms
11.4 What centaurs cannot solve: structural injustices and hard power

12. Conclusion: First People, Then Fields
12.1 Reasserting the ordering: why centaurs must come first
12.2 Ethical centaurs as living checks on technocratic drift
12.3 Directions for research, practice, and institutional innovation
12.4 A caution and an invitation for those who would shape AI futures

13. References
13.1 Work on human–AI collaboration, centaurs, and hybrid intelligence
13.2 Literature on socio-technical systems and large technical infrastructures
13.3 Governance, regulation, and ethics of AI and adjacent technologies
13.4 Empirical and theoretical studies on alignment, agency, and systemic risk

1. Introduction: Reordering the Alignment Conversation

The story of AI alignment is often told as a sequence of expanding technical challenges. First came models: large neural networks whose behavior could be tuned by training procedures and evaluated on benchmarks. Then came stacks: complex assemblies of models, tools, memory, and interfaces, deployed into workflows and products. The natural next step, many argue, is to think in terms of fields: whole AI ecosystems comprising multiple stacks, institutions, infrastructures, and regulatory regimes.

This sequence is seductive because it mirrors the technical and economic expansion of AI itself. As systems scale and entangle, it seems appropriate to move to higher levels of abstraction. Yet there is a risk hidden in this trajectory. If we leap too quickly from stacks to fields, we risk bypassing the most important unit of moral agency in the current transition: the enduring partnership between a human and their AI systems, the ethical centaur.

This paper proposes a deliberate reordering of the alignment conversation. Rather than treating field-level governance as the obvious "next frontier," it argues for the priority of ethical centaurs. Before societies can legitimately design or govern AI fields, they must first cultivate a widespread culture of reflective, accountable human–AI dyads. Without such centaurs, any talk of field governance risks collapsing into technocratic control exercised by those who already command infrastructural and economic power.

The following subsections set up this argument. They trace the conceptual move from models to stacks to fields, explain why "centaurs before fields" is both a moral and practical claim, and outline the scope and structure of the paper.

1.1 From models and stacks to the temptation of fields

The initial phase of contemporary AI was dominated by models. Research, investment, and media attention converged on training larger and more capable systems and on demonstrating their prowess on benchmarks. Alignment, in that context, meant shaping loss functions, datasets, and fine-tuning procedures so that a given model behaved in more helpful, honest, and harmless ways.

As models were embedded into products, it became clear that users never encounter a model "by itself." They encounter stacks: systems that orchestrate calls to multiple models, retrieve information from external sources, maintain state across interactions, and present outputs through interfaces. Misuses and failures often arose not from model internals alone, but from how the stack routed tasks, framed prompts, handled uncertainty, and integrated outputs into decision-making processes.

Recognizing this led to a shift: capability was increasingly understood as a property of stacks, not just of models. Alignment discussions accordingly broadened to include user experience design, workflow integration, and the distribution of responsibility between automation and human oversight.

From there, the idea of fields presents itself almost irresistibly. Once many stacks interact within shared infrastructures, markets, and regulatory environments, the resulting behavior exhibits properties that cannot be reduced to any single component. Concentration of power, systemic risk, informational externalities, and cultural drift all manifest at this larger scale. It seems natural, then, to speak of AI fields and to imagine that alignment must eventually be addressed at that level.

This is the temptation: to see the move to fields as simply the next logical step in a series, and to rush into designing field-level governance architectures while assuming that the underlying human–AI relationships will somehow sort themselves out. The central claim of this paper is that such a move is premature and potentially dangerous.

1.2 Why “centaurs before fields” is a moral and practical claim

The slogan “centaurs before fields” is not merely a call to slow down or to cling nostalgically to small-scale interactions. It makes a substantive claim about where alignment lives and how legitimate governance can emerge.

Morally, alignment concerns the relationship between technical systems and human values. Those values are not abstract invariants; they are interpreted, contested, and enacted by people in specific contexts. Ethical centaurs are the locus where this interpretation and enactment happen in real time. In a centaur, a human does not merely operate a tool or supervise a system; they engage in an ongoing partnership, learning how to delegate, when to override, and how to integrate AI-generated suggestions into their own judgment. Responsibility is felt in the dyad, not in a remote field-level dashboard.

If field-level governance is designed without a critical mass of such centaurs, it will be designed over people rather than with them. Decisions about acceptable trade-offs, risk tolerance, and the distribution of benefits and burdens will be made by a small subset of actors who control infrastructure and policy. The result may be technically sophisticated, but it will lack moral grounding in the lived practice of those who must inhabit the resulting field.

Practically, ethical centaurs are also a precondition for meaningful field-level interventions. Field governance relies on signals: incident reports, feedback, professional norms, and the capacity of individuals and organizations to detect and articulate misalignment. If typical human–AI interactions are shallow, unreflective, or purely instrumental, the field will be starved of high-

quality information about harms and failures. Even well-intentioned regulators and standards bodies will be left navigating by noisy, lagging indicators.

By contrast, when many individuals and teams operate as ethical centaurs, they collectively form a distributed sensor network for alignment. They notice subtle shifts in how AI systems affect their work, relationships, and institutions. They can push back against misuses, articulate domain-specific norms, and contribute to the formation of shared standards. In this way, centaurs do not replace field governance; they anchor it.

“Centaurs before fields” is therefore a statement about sequencing. It does not deny that field-level analysis and policy will be needed. It asserts that, without investing first in centaur formation and culture, field-level moves are likely to harden misalignments rather than correct them.

1.3 Scope, audience, and structure of the paper

This paper is addressed to several overlapping audiences. It speaks to researchers and practitioners in AI safety and alignment who are beginning to grapple with systemic and governance questions. It addresses policymakers and institutional leaders who may be tempted to frame AI primarily as a field-level regulatory challenge. It also speaks to professionals and advanced users who are already living as de facto centaurs but may lack language to describe their role.

The scope is deliberately conceptual and normative rather than narrowly technical. The paper does not propose specific algorithms, protocols, or legal texts. Instead, it offers a reframing: it identifies ethical centaurs as the primary unit of analysis for alignment in the near and medium term, and treats fields as emergent structures whose quality depends on the density and health of such centaurs.

The argument proceeds in several steps. After this introduction, the paper briefly revisits the model–stack–field progression to establish common ground. It then introduces ethical centaurs as human–AI dyads and analyzes their anatomy, norms, and contexts. Subsequent sections examine how networks of centaurs form proto-fields, why fields built without centaurs tend to drift toward control, and how centaur-rich environments can support more legitimate and responsive forms of field governance. The paper concludes with design principles for cultivating ethical centaurs, scenarios contrasting centaur-poor and centaur-rich futures, and a set of open questions for research and institutional innovation.

The aim is not to settle debates about alignment, but to suggest a different center of gravity for them. By insisting that alignment is first a matter of ethical dyads and only then a matter of fields, the paper hopes to reorient both technical and governance efforts toward the human–AI partnerships in which our shared future will actually be negotiated.

2. Background: Models, Stacks, and the Emergence of Fields

Before arguing for "centaurs before fields," it is useful to review how the discourse around AI capability has moved through three natural levels of description: models, stacks, and fields. Each level adds something real. Models give us raw capability, stacks give us situated behavior, and fields give us the large-scale patterns that emerge when many systems and institutions interact. The problem is not that any one of these levels is wrong; the problem arises when attention moves up a level without preserving what mattered at the level below.

This section revisits each step in turn. It starts with models as the scalar unit of capability, then describes stacks as the practical unit of deployment and experience, and finally characterizes fields as distributed environments of interacting stacks and institutions. It ends by highlighting the specific risk that motivates this paper: the temptation to jump directly from stacks to field governance, bypassing the centaur as the primary locus of alignment work.

2.1 Models as the scalar unit of capability

In the first wave of modern machine learning, especially with the rise of large neural networks, the model became the natural unit for thinking about AI. A model was a single, parameterized function that mapped inputs to outputs: text to text, images to labels, states to actions. Capability was primarily described in terms of the model's internal scale and its performance on standardized benchmarks.

This scalar view had several virtues. It made progress legible: more parameters, more data, and better architectures led to clear improvements on tasks like language modeling, image classification, or game playing. It encouraged the development of shared evaluation practices, so that researchers and companies could compare their systems and track state of the art. It made alignment look like a problem of shaping the loss surface and training process so that the model learned safe and useful behaviors.

Within this framing, the central questions were technical: how to avoid overfitting, how to make models robust to adversarial examples, how to reduce toxic or biased outputs by modifying data and objectives. Safety and alignment initiatives focused heavily on the model's internal representations and on interventions during training or fine-tuning. When misbehavior occurred, the default assumption was that something inside the model needed to be changed.

This scalar model-centric era is not over; foundation models still sit at the core of many systems. But as these models were embedded in real products, their limitations as a standalone unit of analysis became increasingly obvious. Users never interact with a raw model. They interact with something more complicated.

2.2 Stacks as the practical unit of deployment and experience

As AI moved from laboratories into applications, organizations discovered that the user's experience of intelligence depended on much more than the behavior of a single model. Systems that felt powerful and reliable were not just "good models"; they were carefully composed stacks.

A stack can be understood as a layered assembly of components: one or more models, retrieval and search tools, structured databases, memory mechanisms, logging and monitoring systems, user interfaces, and integration points with other software. At this level, questions of orchestration become central. Which model is called when. What information is retrieved or cached. How is uncertainty represented to the user. Where is human review mandatory.

From the user's perspective, stacks are the practical unit of capability. Two organizations might use similar base models, but end up with very different systems because their stacks differ. One stack might aggressively automate decisions, while another insists on human-in-the-loop review. One might expose model internals for inspection, while another presents only polished outputs. One might prioritize speed and convenience, another deliberation and documentation.

Alignment discussions therefore shifted to include stack design. It became clear that many harms did not come from models generating obviously dangerous content, but from stacks that routed sensitive tasks to them without proper guardrails, failed to log critical actions, or framed prompts in ways that encouraged overconfident hallucination. Questions of responsibility also shifted: alignment was not only about model weights, but about the product and workflow choices made by designers, engineers, and managers.

Stacks thus became the practical unit of deployment and experience. They are where most current alignment work in industry naturally lives: in red teams, design reviews, human factors studies, and safety testing of end-to-end systems. Yet even stacks are not the whole story.

2.3 Fields as distributed environments of interacting stacks and institutions

When many stacks operate within shared infrastructures, markets, regulatory frameworks, and cultures, they form something larger: an AI field. A field is not a single system but an environment composed of multiple interacting systems, organizations, and rules. It includes cloud and chip infrastructures, data and attention flows, business models, legal regimes, professional norms, and the practices of end users.

At this level, new phenomena appear. Power concentrates around certain infrastructure providers, model developers, or platforms. Externalities accumulate as AI-mediated content reshapes public discourse, labor markets, and institutional trust. Systemic risks emerge as many

organizations become dependent on similar models, data sources, or infrastructure. Cultural patterns arise as people adapt their expectations and behaviors to AI-saturated environments.

Individual models and stacks contribute to these field-level patterns, but none of them determines the outcome alone. A field is the result of many local choices interacting through feedback loops and constraints. It is entirely possible for every stack to meet a plausible set of local safety criteria, while the field as a whole drifts toward outcomes that are misaligned with widely held values: erosion of human skills, consolidation of power, pervasive surveillance, or epistemic fragmentation.

The move to a field lens is therefore not optional. It becomes necessary once AI is deeply embedded into critical infrastructures and social systems. Without it, alignment efforts risk addressing only local behaviors while ignoring the global configuration that those behaviors collectively produce.

2.4 The risk of jumping directly from stacks to field governance

Given this background, it may seem natural to extend alignment work from stacks to fields and to focus on designing governance frameworks at the field level: regulations, standards, treaties, and oversight institutions. The risk, and the motivation for this paper, lies in doing so without first attending to the quality of the individual human–AI relationships that populate the field.

Field governance is inherently political. It involves decisions about who sets standards, who has enforcement power, how trade-offs are made between efficiency and protection, and whose values are privileged in the process. If such governance arises in a context where typical human–AI interaction is shallow, unreflective, or dominated by a small number of providers, then the field lens becomes a tool for consolidating control rather than for aligning technology with the lived values of its users.

Jumping directly from stacks to field governance also creates a diagnostic problem. Field-level interventions depend on information about how AI systems are actually used and what harms they produce. That information is generated first and most vividly in human–AI dyads: in the judgments of professionals, workers, and citizens who experience AI's effects directly. If these dyads are not cultivated as ethical centaurs—capable of noticing misalignment, articulating concerns, and resisting misuse—then field governance will be operating in the dark, guided more by abstract models and political incentives than by grounded feedback.

Finally, there is a risk of moral displacement. When attention moves quickly to fields, individuals and organizations can come to see alignment as someone else's responsibility: a matter for regulators, standards bodies, or international agreements. The concrete work of practicing responsible human–AI collaboration in everyday settings is overshadowed by debates about

high-level governance. The field becomes a stage on which a small number of actors negotiate, while the centaurs who actually live with AI are reduced to subjects rather than partners.

The remainder of this paper is a response to that risk. It accepts the reality and importance of AI fields, but argues that any legitimate field-level governance must emerge from, and remain accountable to, a dense population of ethical centaurs. To make that case, the next sections turn to defining centaurs, analyzing their internal structure, and exploring how networks of such dyads can themselves give rise to more humane and steerable AI fields.

3. Ethical Centaurs: Human–AI Dyads as the Primary Unit

If models are engines and stacks are vehicles, ethical centaurs are drivers who treat driving as more than just getting from point A to point B. They do not simply operate a machine; they cultivate a shared practice with it. In this view, the basic unit of alignment is not "the system" in isolation, but the ongoing relationship between a particular human and the AI systems they work with over time.

This section clarifies what is meant by a centaur in this context, and what distinguishes an ethical centaur from a merely effective one. It then examines how agency, responsibility, and veto power are distributed inside the dyad and contrasts centaur regimes with two common alternatives: fully automated pipelines and fully manual workflows. The goal is to show that centaurs are not a vague metaphor but a concrete, analyzable configuration with distinct alignment properties.

3.1 Defining centaurs: beyond tool use, toward shared practice

The term "centaur" is sometimes used loosely to describe any situation where a human uses an AI system. That is too broad for present purposes. A person running a single query through a search interface is interacting with AI, but they are not yet a centaur in the sense that matters here.

For this paper, a centaur is defined by at least three characteristics.

First, there is continuity of relationship. A centaur is not a one-off interaction but a human and AI system working together repeatedly, in a domain that matters to the human's goals or responsibilities. Over time, the human learns how the system behaves, where it is strong or weak, and how to frame tasks. The AI, in turn, may adapt to the user via preferences, fine-tuning, or memory. The dyad accumulates shared history.

Second, there is a meaningful division of cognitive labor. In a centaur configuration, the human and AI each take on roles that neither could execute as well alone. The AI might handle large-

scale pattern recognition, drafting, or simulation, while the human provides contextual understanding, ethical judgment, and meta-level direction. The key is that the human is not simply delegating everything downstream; they are choosing what to delegate and what to retain.

Third, there is an emerging shared practice. Over time, the centaur develops routines, heuristics, and norms specific to the dyad. These might include agreed ways of checking outputs, habits of asking for alternative options, or internal signals the human uses to recognize when they are over-relying on the system. This practice is not encoded in the model weights or in the human alone; it resides in the recurring interaction between them.

Under this definition, being a centaur is not simply "using a tool" in a casual sense. It is more akin to working with a colleague or instrument that changes how one thinks and acts. The centaur is the human–AI dyad plus its accumulated practice.

3.2 Ethical versus merely effective centaurs

Not all centaurs are ethical. It is possible for a human–AI dyad to be extremely effective at achieving some objectives while being indifferent or hostile to broader ethical concerns. A profitable but manipulative marketing operation, a finely tuned disinformation engine, or an aggressive cost-cutting system that erodes working conditions could all be organized as centaurs if a human and AI collaborate closely and effectively toward those ends.

The distinction between ethical and merely effective centaurs therefore lies not in the presence of collaboration but in the orientation of that collaboration. Several features characterize an ethical centaur.

First, ethical centaurs acknowledge external stakeholders. They do not confine their attention to the dyad's internal goals but consider the effects of their actions on clients, patients, colleagues, or affected communities. This does not mean that they perfectly satisfy all obligations, but that they recognize these obligations as part of their practice.

Second, ethical centaurs maintain explicit norms of reflection and restraint. They ask not only "Can we do this more efficiently" but "Should we do this at all" and "What failure modes should we anticipate." They build routines for second looks, peer consultation, or human-only judgment in high-stakes cases. These norms might be personal, professional, or organizational, but they are integral to how the dyad operates.

Third, ethical centaurs treat the AI's outputs as proposals, not commands. They retain the habit of questioning, cross-checking, and sometimes rejecting the system's suggestions even when those suggestions are usually helpful. This preserves the human's capacity for moral and epistemic independence, rather than allowing it to atrophy.

Fourth, ethical centaurs are attentive to their own drift. They recognize that working with a powerful AI can subtly change their standards, expectations, and risk tolerance. They periodically step back to reassess whether their practice is becoming more aligned with their values or slowly pulled toward objectives they would not endorse if made explicit.

In contrast, a merely effective centaur may optimize ruthlessly for performance metrics, profit, or speed, while neglecting these dimensions of stakeholder consideration, reflection, and self-scrutiny. Both types of centaurs are real. The argument of this paper is that alignment worth pursuing at scale requires the ethical kind.

3.3 Agency, responsibility, and veto power inside the dyad

Ethical centaurs are defined not just by what they achieve but by how agency and responsibility are arranged within the dyad. Three elements are central: who initiates actions, who bears responsibility for outcomes, and who possesses veto power.

Initiation concerns who decides what the AI is asked to do. In many configurations, the human initiates tasks by setting goals, drafting prompts, or selecting data. However, as systems become more proactive—issuing recommendations, surfacing alerts, or proposing workflows—initiation can blur. An ethical centaur remains conscious of this shift and deliberately chooses whether to accept or ignore AI-initiated suggestions. They treat such suggestions as inputs into their agency, not as automatic triggers.

Responsibility, in an ethical centaur, stays anchored in the human. The AI can be seen as an amplifier, assistant, or instrument, but it is not treated as the bearer of blame or praise. When an outcome is harmful, the centaur's human member does not say "the system decided" as if that ended the inquiry. Instead, they ask: why did I accept that output, what assumptions did I make, what checks did I skip, and what changes should I make in the practice going forward. This does not absolve organizations of responsibility for designing and deploying systems, but it preserves an internal stance of ownership.

Veto power is where ethical centaurs differ most sharply from emerging "human in the loop" clichés. In a genuine ethical centaur, the human has practical, not just nominal, veto power over the AI's contributions. This means they have time, institutional permission, and interface support to reject or override the system when their judgment or values conflict with it. If organizational pressures or interface designs make it costly or impossible to exercise this veto, then the centaur is hollow; the human is effectively an operator executing the system's will.

There can also be a reversed form of veto: circumstances in which the system constrains the human from taking certain actions, for example by warning of legal or safety violations or refusing to perform prohibited tasks. In an ethical centaur, such constraints are transparent and

revisable through legitimate processes. They are not silent or opaque limitations that quietly reshape human agency without explanation.

In sum, ethical centaurs preserve and structure human agency inside the dyad rather than dispersing it into the system. Their alignment properties arise from this arrangement of initiation, responsibility, and veto power.

3.4 How centaurs differ from fully automated and fully manual regimes

To understand why centaurs matter, it is useful to contrast them with two limiting cases: fully automated regimes and fully manual regimes.

In a fully automated regime, human involvement is largely confined to initial design and occasional oversight. Once deployed, the system takes inputs, processes them through models and rules, and produces outputs that are executed with minimal human intervention. Automated trading algorithms, content ranking pipelines, or batch decision engines can approximate this pattern.

Fully automated regimes offer clear benefits in speed and consistency, but their alignment properties are fragile. When the world changes or unforeseen cases arise, there may be no human in the immediate loop who both understands the context and feels personally responsible for outcomes. Feedback about harms can be slow or indirect, and there may be strong incentives to ignore it if the system remains profitable or efficient. Moreover, fully automated regimes can erode human skill and situational awareness over time, making it harder to intervene meaningfully when things go wrong.

Fully manual regimes, by contrast, rely on human judgment and labor at every step. Decisions are made without AI support, based on training, experience, and available tools that are not themselves adaptive or generative. Many professional practices historically looked like this: diagnosis without decision support, editing without language models, planning without predictive analytics.

Manual regimes can be aligned when institutions, training, and norms are healthy. They keep humans intimately involved in every decision, preserving skill and responsibility. But they can be slow, inconsistent, and limited in capacity. They struggle with scale and complexity that exceed human cognitive bandwidth, and they may systematically miss patterns that data-intensive methods could reveal.

Ethical centaurs occupy a middle ground with distinctive properties. Compared to full automation, they keep a human in genuine control of high-level goals, interpretations, and vetoes. Compared to fully manual regimes, they augment human capabilities with computational insight and speed. Crucially, they treat the human–AI interaction itself as a site

for alignment work: a place where norms can be negotiated, errors noticed, and practices adjusted.

This does not mean that every process must be organized as a centaur. Some low-stakes or highly standardized tasks may appropriately be fully automated. Some high-stakes or deeply interpersonal tasks may appropriately remain largely manual. The claim is that for a wide class of impactful decisions and creative work, ethical centaurs are the configuration that best balances capability with responsibility.

The next sections build on this foundation. They examine how ethical centaurs operate in specific domains, how networks of such dyads become proto-fields, and why the density and health of centaurs should be treated as a leading indicator for any attempt to align AI at larger scales.

4. Anatomy of an Ethical Centaur

An ethical centaur is not just a human plus a powerful model. It is a structured relationship with a particular shape: a patterned division of labor, a set of shared norms and protocols, a level of mutual transparency, and a rhythm of attention over time. These elements distinguish an ethical centaur from an ad hoc user with a fancy tool. They determine whether the dyad amplifies wisdom or merely accelerates existing biases and blind spots.

This section dissects that structure. It starts with the cognitive division of labor between human and AI, then turns to the norms and protocols that embed reflection and error correction into everyday practice. It then considers what transparency and explainability mean inside the dyad, and closes by examining the temporal dimension: how time and attention are managed so that judgment is cultivated rather than eroded.

4.1 Cognitive division of labor: what the human keeps, what the AI takes

In a naive configuration, the AI is treated as a generic cognitive amplifier: “give it whatever you can, let it do as much as possible.” Ethical centaurs are more deliberate. They make explicit decisions about what kinds of cognitive work belong with the human and what kinds can safely and beneficially be delegated.

Typically, the AI takes on tasks that involve wide search, pattern recognition across large datasets, rapid generation of variants, and mechanical consistency. It drafts, summarizes, explores alternatives, simulates scenarios, and checks formal constraints. These are domains where computational scale and speed excel and where errors, if properly checked, can be identified and controlled.

The human, in contrast, keeps tasks that involve value judgments, interpretation of ambiguous context, understanding of institutional and interpersonal nuance, and responsibility for final commitments. The human decides which questions to ask, which options are even worth considering, how to frame trade-offs, and when to slow down or stop. This is not because humans are magically infallible, but because they are embedded in moral, social, and legal networks in ways that current systems are not.

Importantly, this division of labor is not fixed. As the centaur matures, the human may discover that some tasks initially reserved for themselves can be safely delegated, provided appropriate checks exist. Conversely, they may reclaim tasks when they discover subtle harms from delegation. The division is a living contract. Ethical centaurs revisit it periodically rather than allowing the boundary to drift silently toward "the AI does everything and I just approve."

4.2 Norms and protocols: reflection, doubt, and error correction

A centaur becomes ethical not only by what it does but by how it handles uncertainty and failure. Norms and protocols for reflection, doubt, and error correction are therefore central to its anatomy.

Reflection norms govern how the dyad pauses and thinks about its own process. For example, before acting on a recommendation in a high-stakes case, the centaur might require a second round of questioning: "What assumptions is this suggestion making. What data might be missing. What would it take for this to be wrong." These questions can be posed by the human, by the AI if designed to self-criticize, or by both in dialogue. The key is that some part of the dyad is tasked with meta-level evaluation, not just first-pass output.

Doubt is treated as a feature, not a bug. Ethical centaurs cultivate habits of justified skepticism, especially when an output is unusually convenient, flattering, or aligned with existing biases. They learn to recognize telltale signs that the system may be extrapolating beyond its training or misreading context. Doubt does not mean paralysis; it means selectively increasing scrutiny when the situation calls for it.

Error correction is organized, not ad hoc. When mistakes are discovered—whether factual, ethical, or relational—the centaur does more than fix the immediate case. It asks what went wrong in the practice. Was the prompt misleading. Were key constraints omitted. Did time pressure override a check that should have been mandatory. Ethical centaurs build small but robust protocols for capturing such lessons, whether as updated prompts, checklists, or "red flag" conditions.

These norms and protocols are lightweight but intentional. They cannot eliminate error, but they change how errors propagate. Instead of errors accumulating silently, they become

occasions for tightening the practice. Over time, the centaur's shared routines become a form of embodied alignment: a way of ensuring that power is coupled to conscientiousness.

4.3 Transparency and explainability within the dyad

Discussions of AI transparency often focus on external observers: regulators, auditors, or the public. Ethical centaurs require a more intimate kind of transparency: the ability of the human to understand what the AI is doing well enough to cooperate and to contest.

At minimum, this means that the AI's outputs are accompanied by cues that help the human assess reliability. These might include indications of uncertainty, references to sources or constraints, or alternative options ranked by confidence. The point is not to fully expose internal weights or training data, but to give the human hooks they can use to interrogate the output: "Why this. What else could you have said. What did you ignore."

Transparency also runs the other way. The AI needs enough visibility into the human's goals, constraints, and context to avoid naive inference. Ethical centaurs achieve this not by asking the user to expose everything, but by developing a shared vocabulary for specifying what matters: "This is high-stakes and irreversible," "This concerns a vulnerable person," "This is exploratory and low-risk." These tags, formal or informal, guide the system's behavior and prompt it to request clarification rather than guessing.

Explainability, inside the dyad, is less about producing polished rationalizations for external consumption and more about supporting genuine dialogue. When the human asks, "Why did you recommend this option," the system should be able to surface relevant patterns, not just generic statements. When the system flags a human decision as inconsistent with past practice, it should be able to point to the discrepancy in a way that invites conversation rather than passive compliance.

In a healthy centaur, this mutual transparency builds trust, but not blind trust. The human comes to know when the system is usually reliable and when it tends to err. The system, in turn, adapts to the human's preferences and risk tolerances. The result is a relationship in which each party can question the other with some hope of being heard and understood.

4.4 Time, attention, and the cultivation of judgment

Ethical centaurs exist in time. Their practices depend on how time and attention are allocated, both within individual decisions and across longer arcs of collaboration. Without attention, even the best norms and interfaces degrade into rituals or are bypassed under pressure.

On the short timescale, the centaur must decide when to move fast and when to slow down. Not every action can be preceded by elaborate reflection; some decisions are routine. Ethical centaurs therefore stratify their workflows: low-stakes, reversible tasks can be processed quickly

with light checks, while high-stakes, irreversible, or ethically charged tasks trigger slower, more deliberative modes. Interfaces and organizational expectations must support this stratification; if everything is rushed, the human veto becomes symbolic.

On the medium timescale, attention is needed for review. Ethical centaurs allocate time to look back over sequences of decisions, patterns of reliance on the AI, and emerging anomalies. This can take the form of periodic debriefs, case reviews, or simply personal reflection. Without such review, drift accumulates unnoticed: thresholds slide, shortcuts become normal, and initial intentions are forgotten.

On the long timescale, judgment is cultivated. Working with an AI system can either erode or deepen the human's capacity for discernment. If the centaur's practice treats the system as an oracle and the human as an executor, the human's skills will atrophy. If, instead, the practice routinely asks the human to interpret, to compare alternatives, to explain decisions to others, and to bear responsibility, then the AI becomes a tutor in judgment as well as a tool.

Attention is also needed to protect the centaur from manipulation. Business models and interface designs may push toward constant engagement, maximizing time spent in AI-mediated environments. Ethical centaurs resist becoming always-on channels for external incentives. They preserve spaces and times where the human steps away from the dyad to think, rest, and be influenced by other sources: colleagues, texts, lived experience.

In this way, the anatomy of an ethical centaur is as much temporal as structural. It is about creating rhythms in which fast, powerful computation is woven into slower, reflective, human time. Those rhythms, more than any specific prompt or parameter, determine whether the centaur sustains alignment or quietly trades it away for speed and convenience.

5. Centaurs in Context: Work, Professions, and Everyday Life

Ethical centaurs are not an abstract ideal; they are already emerging in many domains wherever humans rely on AI systems in sustained, consequential ways. The shape of the centaur varies with context. A physician co-diagnosing with a model, a legal researcher drafting with a large language system, an engineer exploring design spaces, a musician co-composing, a middle manager triaging information, and an ordinary person managing their day with an assistant all inhabit different constraints and expectations.

Understanding these contextual differences matters for alignment. The same human–AI configuration that is ethically promising in one domain can be dangerous in another. Professional codes, regulatory regimes, organizational incentives, and power asymmetries all influence whether a centaur is likely to be ethical, merely effective, or misaligned. This section

sketches how centaurs appear in expert domains, creative and cultural work, administrative and managerial roles, and everyday life, highlighting both opportunities and failure modes.

5.1 Centaurs in expert domains: medicine, law, science, engineering

Expert domains already have dense infrastructures of responsibility: professional training, licensing, malpractice regimes, peer review, and ethical codes. These can act as scaffolding for ethical centaurs or, if misused, as camouflage for over-delegation.

In medicine, a centaur might consist of a clinician and a diagnostic or triage system. The AI system can surface differential diagnoses, highlight rare conditions, suggest treatment guidelines, or flag potential drug interactions. The clinician contributes context about the patient's history, values, social situation, and non-verbal cues, and makes final decisions about consent and trade-offs. An ethical medical centaur maintains the clinician's responsibility for the patient's well-being, uses AI primarily to expand and refine human judgment, and incorporates systematic checks when recommendations conflict with experience or guidelines. Misalignment arises if time pressure, billing incentives, or institutional policies push clinicians to accept AI suggestions without critical scrutiny or to offload conversation and empathy onto systems.

In law, centaurs may appear in research, drafting, and strategy. AI systems can rapidly surface relevant precedents, propose argument structures, or simulate how different fact patterns might fare in different jurisdictions. An ethical legal centaur preserves the attorney's responsibility to understand and argue the case, to safeguard client interests, and to respect procedural fairness. The lawyer checks citations, tests arguments for coherence, and resists the temptation to let AI-generated text stand without review. Misalignment occurs when AI systems are treated as black-box authorities, when ghostwritten documents erode accountability, or when powerful tools are used to overwhelm less resourced parties with volume rather than substance.

In science, centaurs can accelerate hypothesis generation, literature review, data analysis, and even experimental design. A scientific centaur might use AI to scan vast literatures for overlooked connections, propose candidate models, or detect subtle patterns in data. The human scientist contributes domain intuition, skepticism, and the design of robust tests. Ethical scientific centaurs remain anchored in reproducibility, transparency about AI assistance, and humility about uncertain results. They guard against overfitting to patterns discovered by systems whose inductive biases are poorly understood. Misalignment surfaces when AI is used to manufacture plausible but shallow papers, fabricate data, or chase fashionable metrics disconnected from inquiry.

In engineering, centaurs may co-design circuits, software architectures, or mechanical structures. AI tools can explore design spaces, suggest optimizations, or anticipate failure modes

under varied conditions. The engineer understands constraints, safety margins, regulatory requirements, and long-term maintainability. Ethical engineering centaurs treat AI suggestions as candidates to be justified, simulated, and tested, not as guarantees. They remain aware that small design decisions, magnified at scale, can have significant safety and environmental consequences. Misalignment arises when cost and speed pressures lead to uncritical acceptance of AI-optimized designs without sufficient robustness analysis.

Across these domains, the presence of professional norms and external accountability mechanisms creates fertile ground for ethical centaurs. But these same structures can be co-opted to rationalize over-reliance. Whether centaurs become genuinely ethical depends on how deeply human practitioners integrate AI into their understanding of responsibility, not just their workflow.

5.2 Centaurs in creative and cultural work

Creative fields present a different configuration. Here, the stakes are often framed less in terms of immediate physical harm and more in terms of meaning, identity, and culture. Centaurs in these domains grapple with questions of authorship, originality, and artistic integrity as they collaborate with generative systems.

For writers, AI tools can suggest plotlines, generate dialogue, or propose alternative phrasings. A writing centaur might use these capabilities to break creative blocks, explore different voices, or rapidly iterate on drafts. Ethical practice involves more than avoiding plagiarism; it includes transparency about the role of AI, reflection on how reliance on generative systems shapes one's style, and attention to whose voices and perspectives are amplified or suppressed by training data. A merely effective centaur might produce large volumes of plausible text optimized for engagement, while an ethical centaur uses AI to deepen craft and expand imagination, not to flood discourse with generic output.

Visual artists and designers encounter centaurs in tools that generate images, layouts, or stylistic variations. AI can help explore visual concepts, translate rough sketches into refined renderings, or experiment with combinations of styles. Ethical centaurs in this space consider how model training on existing artworks implicates questions of appropriation and consent, how their use of AI affects the livelihoods of other creators, and how the aesthetics they propagate influence cultural norms. They may choose to constrain their tools, to opt for models trained on licensed or self-generated data, or to foreground process in how works are presented.

Musicians and composers can work as centaurs with systems that suggest chord progressions, harmonies, melodic continuations, or production choices. These tools can act as improvisational partners or scaffolding for exploring new genres. Ethical centaurs in music retain a sense of

voice and intentionality, using AI as a collaborator rather than as a generator of background noise for maximal consumption. They are aware of the risk that commodified generative systems can saturate cultural spaces with homogenous sound, crowding out experimental or minority traditions.

In all these creative domains, centaurs are not just optimizing for accuracy or efficiency; they are participating in the evolution of culture. Ethical practice acknowledges this. It resists the reduction of creativity to content production and treats AI as a medium that requires its own literacy and ethics. Misaligned centaurs in creative work may contribute to a cultural field characterized by homogenization, superficiality, or exploitation, even if no single piece of work appears harmful in isolation.

5.3 Centaurs in administrative, managerial, and bureaucratic roles

Administrative, managerial, and bureaucratic roles often involve processing information, making medium-stakes decisions, and enforcing policies across large populations. In these contexts, centaurs can either humanize systems by enabling better understanding and responsiveness, or further depersonalize them by making large-scale surveillance and control more efficient.

Administrative centaurs might use AI to sort and prioritize communications, schedule resources, or match applications to programs. Ethical centaurs in these roles remain attentive to fairness, transparency, and the impact of errors on individuals. They use AI to surface cases that need human attention, not to automate away judgment. They understand that a misclassified loan application, benefit claim, or complaint is not an abstract data point but a real disruption in someone's life.

Managers may work as centaurs with tools that analyze team performance, predict attrition, or recommend interventions. Ethical managerial centaurs resist the temptation to treat workers as data points to be optimized. They use AI to gain insight into structural issues, to identify where support is needed, and to check their own biases, rather than to justify intrusive monitoring or punitive micro-optimization. They remain willing to override system recommendations when they conflict with humane treatment or contextual understanding.

In bureaucratic and regulatory settings, centaurs could arise where civil servants use AI to analyze policy outcomes, detect irregularities, or draft regulations. Ethical centaurs here are mindful that AI systems embed assumptions about what counts as a problem, which metrics matter, and whose interests are prioritized. They treat AI outputs as starting points for deliberation with stakeholders rather than as neutral reflections of reality.

The risk in these roles is that centaurs become instruments of distant priorities. When organizational incentives emphasize throughput, cost reduction, or metric improvement above all, the human side of the dyad can be pressured into rubber-stamping system outputs. Veto

power becomes nominal. Ethical centaurs in administrative and managerial contexts thus require institutional backing: policies that protect time for review, that prohibit blind automation in certain decisions, and that recognize conscientious objection to misaligned uses of AI.

5.4 Everyday centaurs: ordinary users, personal assistants, and domestic life

Beyond professional settings lies the vast domain of everyday life, where ordinary users interact with AI through personal assistants, recommendation systems, home devices, and mobile applications. Here, centaurs are often informal and unrecognized. A person who relies on an assistant for scheduling, navigation, message drafting, and information retrieval is, in effect, forming a centaur, even if they would never use that term.

Everyday centaurs differ from professional ones in at least three ways. First, they operate with less explicit training and fewer external norms. There is no licensing body for personal assistant use, no formal code of ethics for everyday navigational choices or entertainment consumption. Second, the power asymmetry between user and provider is typically larger. The assistant's design, data collection, and business model are controlled by remote actors, while the user has limited visibility or leverage. Third, the line between assistance and manipulation is more fluid; systems designed to help can also be optimized for engagement, upselling, or subtle behavioral nudges.

Ethical everyday centaurs therefore depend heavily on defaults. If assistants are designed to be transparent about their limitations, respectful of privacy, and conservative in their attempts to shape user behavior, then informal centaurs can form that genuinely support autonomy. Users can learn to ask clarifying questions, to cross-check important information, and to set boundaries around what tasks they delegate.

However, if everyday assistant systems are optimized primarily for data extraction and attention capture, centaurs in domestic life can become channels through which external incentives reshape habits, relationships, and self-understanding. Children growing up with such systems may come to see constant algorithmic mediation as natural and benign, making it harder to develop independent judgment later.

In domestic contexts, centaurs also mediate shared life: families planning schedules, caring for dependents, or making financial decisions. Ethical centaurs in these situations require awareness not just by individuals but by households. Who controls the assistant's settings. Whose preferences are prioritized. How are conflicts between family members' uses and values resolved. These are subtle but important alignment questions.

Across work, professions, and everyday life, ethical centaurs emerge when humans treat AI systems not as infallible authorities or invisible infrastructure, but as powerful partners whose

use demands reflection and care. The next section considers how such centaurs, once present in large numbers, begin to form networks and proto-fields—and why their density and health should be treated as a prerequisite for any serious attempt at field-level governance.

6. Centaur Networks as Proto-Fields

Ethical centaurs do not exist in isolation. A single human–AI dyad can develop good practices, but the most significant effects emerge when many such dyads begin to interact, observe each other, and share tools, stories, and norms. At that point, something like a proto-field appears: not yet a full-fledged AI field with infrastructures and regulations, but more than a scattered collection of individuals.

These centaur networks are important for two reasons. First, they are where alignment practices are refined and transmitted. Second, they are early sites where power begins to aggregate: around particular tools, prompts, workflows, and gatekeepers of knowledge. Understanding how centaur networks form, how their local micro-institutions arise, and when they cross thresholds into proto-fields is therefore essential for any attempt to shape AI development without sliding prematurely into top-down field governance.

6.1 How centaurs learn from each other and form communities of practice

Centaurs learn not only from their own experience but from each other. As more people work in sustained partnerships with AI systems, they naturally compare notes: what prompts yield reliable behavior, which workflows reduce errors, how to explain to colleagues and clients what the AI is doing, and where it tends to fail.

This exchange of tacit knowledge gives rise to communities of practice. In professional settings, such communities might organize around roles or domains: clinicians using decision support, lawyers using drafting systems, teachers using tutoring tools, engineers using design assistants. In less formal contexts, they might crystallize in online forums, social media groups, or internal chat channels where centaurs share tips, warnings, and anecdotes.

Within these communities, a form of peer review emerges. Stories of near misses and failures spread, prompting others to adjust their own practices. Innovative uses of AI are imitated and adapted. Rules of thumb become common knowledge: “Never trust the system on this kind of case; always double-check that kind of citation; ask for three alternatives before choosing one; do not let it decide this class of issue without a human conversation.”

Over time, these shared practices take on a life beyond any one dyad. New centaurs entering the domain are trained not only in the technical operation of the tools but in the culture of their use. They inherit a repertoire of habits, questions, and red lines. In this way, centaur networks

become channels through which ethical standards can spread horizontally, rather than only from regulators or corporate policies.

At the same time, communities of practice can also propagate misaligned norms. If influential centaurs celebrate speed and volume above all, or normalize cutting corners, those values can spread just as easily. The emergence of centaur networks is thus not automatically beneficial. It is a structural fact: once there are many centaurs, they will form communities. Whether those communities trend toward ethical practice depends on their internal incentives, status structures, and feedback mechanisms.

6.2 Local norms and shared prompts as emergent micro-institutions

As centaur communities mature, they begin to generate artifacts that function as micro-institutions: shared prompts, templates, checklists, style guides, and playbooks. These are not law or formal policy, but in practice they govern behavior.

Shared prompts are a particularly salient example. Within a given professional community, certain prompt formulations can become de facto standards: how to ask for a particular type of analysis, how to instruct the AI to explain its reasoning, how to mark a task as high-stakes or sensitive. These prompts encapsulate collective learning about how to get the system to behave in ways that fit the community's expectations.

Checklists and protocols similarly emerge. A medical centaur community might adopt a simple rule such as "For any AI-suggested diagnosis in category X, verify against at least one independent source and document the reasoning." A legal centaur community might adopt shared guidelines on how to verify AI-suggested citations. In creative communities, shared norms might develop about how much AI-generated material can be included before a work ceases to be considered primarily authored.

These artifacts have institutional properties. They are durable across individuals, they reduce the need to reinvent practice from scratch, and they provide a basis for critique and revision. They can also be contested: different factions within a community may advocate different prompt conventions or protocols, and debates over these can serve as a vehicle for ethical reflection.

The institutional character of these micro-structures becomes more evident when organizations or platforms begin to encode them more formally. For example, a company might bake shared prompts and workflows into its internal tools, or a professional association might issue guidelines based on community-developed practices. At that point, norms that arose bottom-up from centaur practice begin to influence top-down policy.

Ethically, this is a double-edged development. On one side, micro-institutions can stabilize good practices and make it easier for new centaurs to start from a safer baseline. On the other, if problematic norms become codified early, they can be hard to dislodge. The formation of micro-institutions is therefore a key lever point: a stage at which centaur networks can either lock in shallow efficiencies or anchor deeper commitments to reflection and responsibility.

6.3 Centaur clusters in organizations and sectors

Within organizations, centaurs rarely appear at random. They tend to cluster in roles and teams where the combination of high cognitive load and openness to technology adoption is strongest. Over time, such clusters can become influential, shaping how the organization as a whole understands and uses AI.

An organization might have a cluster of centaurs in its data science group, another in legal or compliance, and yet another among early-adopter managers. Each cluster develops its own domain-specific practices, but there is also cross-pollination: techniques and norms developed in one area may be adapted to another.

When organizational leadership is attentive, these clusters can be recognized and supported. Training programs can be designed to spread centaur literacy, internal forums can be established to share practices, and formal roles can be created for people who act as translators between centaur communities and executive decision-makers. In such cases, centaur clusters can help steer the organization toward more thoughtful and accountable AI use.

In other cases, centaur clusters may remain semi-invisible pockets of innovation and risk. They may be tolerated as long as they improve performance, but their ethical concerns or calls for guardrails may be ignored. Alternatively, management may attempt to scale their practices aggressively without preserving the attention and veto structures that made them ethical in the first place, turning living practices into brittle procedures.

At the sector level, centaur clusters can become loci of influence as well. A particular hospital system, law firm, research lab, or creative studio might be seen as a leader in centaur practice, attracting talent and setting informal standards that others follow. Sectoral conferences, journals, and online spaces then become channels through which centaur norms and techniques diffuse beyond organizational boundaries.

These clustering dynamics mean that centaur networks are not spread evenly. Certain organizations and sectors may become dense hubs of centaur practice, while others lag far behind or adopt AI in ways that minimize human agency. As these patterns develop, they begin to resemble the uneven landscape of a proto-field: with nodes of innovation, regions of risk, and emerging structures of prestige and authority.

6.4 When proto-fields become visible: indicators and thresholds

At what point do centaur networks and clusters become more than scattered phenomena and take on the character of a proto-field. There is no single switch, but several indicators suggest that a threshold has been crossed.

One indicator is the emergence of shared narratives. When centaurs across organizations and roles begin to tell similar stories about what it means to work responsibly with AI, what typical failure modes look like, and what good practice requires, a field-like discourse is forming. These narratives may be captured in articles, talks, case studies, or informal lore, but they provide a common frame of reference.

Another indicator is the development of cross-cutting infrastructure specifically aimed at supporting centaurs. This might include specialized tools for logging and reviewing human–AI decisions, training programs focused on centaur skills rather than generic software instruction, or certification schemes that recognize individuals and organizations as practicing a certain standard of centaur ethics. The existence of such infrastructure signals that centaur practice is no longer an incidental adaptation but a recognized domain of expertise.

A third indicator is the appearance of boundary disputes. When questions arise about who has the authority to define good centaur practice, which communities' norms should be privileged, or how conflicts between different centaur cultures should be resolved, we are in the realm of proto-field politics. These disputes might play out within professions, between companies and regulators, or across national contexts.

Finally, quantitative signs can help. One could track, for instance, the proportion of key decisions in a domain that are made in centaur configurations rather than manually or fully automatically; the density of knowledge-sharing channels focused on centaur practice; or the degree to which formal policies reference human–AI dyads as units of responsibility. As these numbers rise, it becomes more appropriate to speak of a centaur-rich proto-field.

Recognizing these thresholds matters for the broader argument of this paper. If centaur networks have not yet solidified into proto-fields, attempts at field-level governance are likely to be disconnected from lived practice. Conversely, once proto-fields are visible and populated by centaurs who have developed their own norms and micro-institutions, they become potential partners in shaping any larger field-level structures.

In this sense, centaur networks are both a warning and an opportunity. They warn that misaligned practices can spread and harden long before formal governance catches up. They offer an opportunity because they are the earliest points at which alignment can be anchored in concrete, everyday human–AI relationships rather than abstract models of behavior. The next sections build on this, arguing that fields built without attending to these proto-structures will

tend toward control, while fields that grow out of centaur networks have a better chance of remaining humane and corrigible.

7. Why Fields Without Centaurs Drift Toward Control

If ethical centaurs are absent or marginal, AI fields do not remain neutral. They evolve under the influence of whoever controls infrastructure, capital, and formal authority. In such conditions, field-level thinking tends to become a language for optimizing populations and systems from above rather than a framework for supporting grounded, distributed agency. The result is a drift toward technocratic control: decisions justified by systemic risk, efficiency, or security, but experienced by most people as opaque imposition.

This section makes that drift more explicit. It examines how technocratic governance becomes a default attractor when centaurs are weak, how concentration of infrastructural and economic power amplifies that tendency, how metric-centric design displaces person-centered practice, and how historical analogies from large technical systems illuminate the dangers of building fields without distributed agency.

7.1 Technocratic field governance as a default attractor

Field-level governance is tempting because it promises a vantage point "above the fray." From that vantage point, policymakers, corporate strategists, and technical experts can talk in terms of systems, flows, and aggregate outcomes. They can define goals such as reducing risk, improving productivity, or protecting national security and then design interventions—standards, regulations, platform policies, infrastructure investments—meant to steer the field toward those goals.

In the absence of strong, articulate centaur communities, these conversations are dominated by actors who already possess expertise, resources, and institutional power. They have the time, information, and channels to participate in standard-setting and policy debates. Everyday human–AI practitioners, especially those outside elite organizations, are largely invisible at this scale. Their practices and concerns are reduced to variables in models or anecdotal inputs into consultation processes.

Technocracy emerges as a default attractor because it solves a real problem: complex fields need some coordination. But when the coordinating agents are insulated from the lived realities of centaur practice, they naturally gravitate toward tools they understand and can justify: quantitative risk models, high-level metrics, and control mechanisms that can be implemented through existing institutions.

Without centaurs pushing back, the logic becomes self-reinforcing. A small group of field managers sets rules that shape how AI can be built and used. Those rules, in turn, structure which kinds of centaur practice can even emerge. Practices that do not fit the prescribed categories are squeezed out or rendered illegible. Over time, the field begins to look less like a diverse ecosystem of human–AI relationships and more like a managed grid of regulated flows.

This is not necessarily the result of malice. It can arise from sincere efforts to "take responsibility" for a powerful technology. The problem is that the field is being governed over the heads of most of the people who live in it. Without centaurs as recognized interlocutors, field-level governance drifts toward control simply because that is the only shape of action available to those at the top.

7.2 Concentration of infrastructural and economic power

The drift toward control is amplified by the structure of AI infrastructure. Building and operating large-scale models, data centers, and cloud platforms requires significant capital and technical capability. As a result, a small number of firms and, in some cases, state entities, control key chokepoints: compute, data pipelines, model distribution channels, and core software stacks.

When such concentration exists, field-level governance often takes the form of negotiations among these powerful entities and regulators. The voices of centaurs—workers, professionals, small organizations, and ordinary users—are mediated through these larger players, if they are heard at all. Even when consultation mechanisms exist, the agenda is shaped by infrastructural and economic realities: what can be monitored, what can be enforced, what aligns with existing business models and strategic interests.

In this environment, the concept of "field alignment" can be co-opted. It may be framed as ensuring that AI development and deployment support certain macro objectives: competitiveness, stability, growth, or security. These are not illegitimate goals, but they are not the same as alignment with the values and experiences of the people who interact with AI daily. If those people are not organized as centaur communities with some degree of collective voice, their interests are easy to overlook or reinterpret.

Moreover, concentrated providers can shape the very tools that centaurs use. Interface designs, default workflows, and built-in guardrails reflect the provider's priorities. If the provider's main alignment concern is liability or public perception, for example, tools may be optimized to reduce visible incident rates rather than to support deep centaur reflection. Users are nudged toward patterns that make the field more legible and manageable from the center, not necessarily more humane on the ground.

This dynamic creates a feedback loop: concentrated power shapes centaur practice, which in turn reinforces dependence on centralized infrastructures. Without alternative infrastructures

or strong countervailing institutions, it becomes increasingly difficult for centaurs to develop and sustain practices that diverge from the dominant provider's vision. Fields without centaurs, or with only weak and fragmented ones, thus naturally align with the interests of those who control the pipes.

7.3 Fields designed around metrics, not persons

Another driver of control-oriented drift is the reliance on metrics as primary guides for field-level decisions. Metrics are indispensable; without them, governance risks becoming arbitrary or purely rhetorical. But when metrics are not grounded in the lived practice of centaurs, they tend to abstract away precisely the aspects of human–AI interaction that matter most.

Field-level metrics might include counts of incidents, measures of model accuracy on benchmark suites, rates of adoption in different sectors, or economic indicators such as productivity gains. These quantities are relatively easy to track and compare. They lend themselves to dashboards and targets. Unfortunately, they say little about whether centaurs feel able to exercise veto power, whether their judgment is being sharpened or dulled, whether they experience their tools as supportive or coercive, or whether marginalized groups are bearing disproportionate burdens.

In a field organized around metrics, interventions are designed to move the numbers. If incident rates are too high, tighten content filters. If adoption is lagging, subsidize deployment. If productivity gains are slow, encourage automation. These moves may be justifiable in narrow terms, but they can have corrosive side effects on centaur practice: encouraging reliance on black-box safeguards instead of thoughtful checking, pushing humans out of loops where their presence would support responsibility, or incentivizing forms of use that maximize measurable outputs at the expense of unquantified harms.

Ethical centaurs can act as counterweights to metric drift. They can articulate misalignments that are not captured in dashboards, such as erosion of trust, loss of skill, or subtle changes in institutional culture. They can insist on qualitative criteria for success: "We consider this deployment acceptable only if practitioners report that they understand and can contest system decisions." Without such voices, however, the field's design defaults to what can be counted.

Over time, metric-centric governance can reshape even people's self-understanding. Practitioners may come to view their worth primarily in terms of the numbers they help generate or control. Human agency is reframed as parameter tuning in larger optimization processes. The field becomes a machine for producing target metrics, and individuals become components, not partners. This is a classic pattern of control, and it is particularly likely to appear in fields where centaur perspectives have not been institutionalized.

7.4 Historical analogies: large technical systems without distributed agency

History offers multiple examples of large technical systems that evolved without strong, distributed agency among those who depended on them. These analogies are not perfect, but they are instructive.

Consider centralized power grids. In many contexts, decisions about generation, distribution, and pricing were made by a combination of state agencies, utilities, and industrial actors. Ordinary users had little say beyond their individual consumption choices. Safety and reliability standards were designed by engineers and regulators, often with genuine public-spirited intent. Yet the system's structure made certain outcomes likely: top-down planning, vulnerability to catastrophic failures, and difficulty integrating alternative energy sources and community-level priorities.

Or consider mass media in the broadcast era. A small number of organizations controlled the channels through which information and culture flowed. Regulation focused on content standards, market structure, and technical parameters. Audience members could adopt, resist, or tune out, but they were not recognized as partners in shaping the field. As a result, systemic biases and manipulative practices could persist for long periods, constrained mainly by competitive pressures and occasionally by public outcry.

Even in domains like industrial production or transportation, large technical systems often developed in ways that subordinated workers' and communities' agency to centralized decision-making. Safety regimes tended to prioritize aggregate risk reduction over participatory consent. Efficiency improvements were frequently pursued without corresponding investments in worker autonomy or voice.

These histories share a pattern: when complex systems are built and governed primarily by a small group of technical and institutional experts, without robust mechanisms for distributed participation, they tend toward control. They may deliver significant benefits, but they also produce externalities, entrench power, and make it difficult for those embedded in the system to steer its evolution.

AI fields risk repeating this pattern in a more acute form. Their flexibility and reach allow them to mediate not just power and information but cognition itself. If they are designed and governed without centaurs—without recognized roles for reflective human–AI dyads as sources of insight and resistance—then they are likely to become even more tightly controlled and less corrigible than earlier infrastructures.

The analogy is not meant to support fatalism. It is a warning that the path of least resistance for large technical systems is toward centralized control, even when the initial intentions are benign. The presence of ethical centaurs, organized into networks and proto-fields, offers one

route to a different outcome: systems in which governance is not only about moving levers at the top, but also about listening to and empowering those who live daily inside human–AI relationships.

The next section turns to that more hopeful possibility, arguing that ethical centaurs are not just desirable micro-level configurations but necessary preconditions for any legitimate and effective form of field governance.

8. Ethical Centaurs as a Precondition for Legitimate Field Governance

If fields tend to drift toward control in the absence of grounded agency, then the presence of ethical centaurs is not a cosmetic improvement; it is a precondition for any form of field governance that can claim legitimacy. A field in which human–AI dyads are shallow, coerced, or absent can certainly be governed, in the narrow sense that someone can set rules and enforce them. But such governance will rest on a thin base of lived understanding and will be prone to both overreach and blind spots.

This section develops that claim. It first sketches what it means for a field to be genuinely populated by ethical centaurs, then considers how such centaurs can be represented in standards, audits, and policy. It contrasts governance over populations with governance with centaurs as partners, and closes by proposing criteria for when field-level interventions become appropriate rather than premature.

8.1 What it means for a field to be populated by ethical centaurs

A field populated by ethical centaurs is not one in which everyone uses AI, nor one in which every AI-using relationship is perfect. Instead, it has several recognizable features.

First, ethical centaurs are common in roles where consequential decisions are made. In medicine, law, engineering, management, and public service, the typical pattern is not full automation or unassisted human labor but reflective human–AI dyads. Practitioners know their tools well, retain responsibility for outcomes, and have real veto power. They are not occasional users dabbling with a novelty interface; they are professionals who have integrated AI into their craft without surrendering judgment.

Second, centaur literacy is widespread. Training and education programs in relevant fields treat human–AI collaboration as a core competency. Students and early-career workers learn not just how to operate tools, but how to think with them, question them, and set boundaries. Ethical considerations are integrated into this training: when to delegate, when to refuse, and how to explain AI-assisted decisions to others.

Third, centaur practice is supported by institutional structures. Organizations allocate time for review, encourage documentation of human–AI decision processes, and protect workers who exercise their veto against misaligned uses. Professional bodies issue guidance based on the accumulated experience of centaurs, not only on abstract principles or vendor claims. There are channels for reporting concerns about harmful patterns in AI use, and these channels are taken seriously.

Fourth, centaurs talk to each other. Communities of practice, as described earlier, are healthy and active. They share not only tricks for efficiency but reflections on the ethical and psychological impact of working with AI. Their norms influence hiring, promotion, and recognition; being an ethical centaur is valued, not treated as an optional personal quirk.

In such a field, ethical centaurs form an epistemic backbone: a distributed network of people who can notice when field-level policies are misaligned with ground reality, articulate the problem in a way others can understand, and propose corrections. Without that backbone, field governance has little to stand on beyond models, metrics, and political bargaining.

8.2 Representation and voice: centaurs in standards, audits, and policy

For ethical centaurs to serve as a foundation for legitimate field governance, they must have more than informal presence; they must have representation and voice in the processes that set standards, conduct audits, and shape policy.

In standard-setting, this means that committees and working groups include not only vendors, regulators, and academic experts, but also people who spend significant portions of their working lives as centaurs in the relevant domain. These practitioners bring concrete knowledge of where AI helps, where it fails, and how it interacts with existing norms and responsibilities. Their role is not merely symbolic; their experiences and judgments must shape definitions, requirements, and best-practice recommendations.

Audits, similarly, should incorporate centaur perspectives. Technical evaluations of models and systems remain important, but they must be complemented by assessments of how human–AI dyads actually function in situ. Do practitioners feel rushed or coerced. Do they understand when and why the system may be wrong. Are they able to override or question outputs without penalty. Audits that ignore these questions risk certifying systems that look aligned in isolation but distort practice when deployed.

In policy-making, centaurs should have organized channels for input. This may take the form of professional associations advocating on behalf of their members, civic organizations representing affected communities, or dedicated consultative bodies composed of practitioners from diverse sectors. The key is that those who live with AI daily are not relegated to sporadic

public comment periods while most of the negotiation occurs between large firms and government agencies.

Representation is not a guarantee of good outcomes. Centaur communities can be captured, fragmented, or unrepresentative of those most affected by AI. But without structured voice for centaurs, field governance is almost guaranteed to tilt toward technocratic and corporate interests. Including centaurs in standards, audits, and policy processes makes it more likely that the field's rules will be responsive to lived practice and more corrigible when they interact badly with that practice.

8.3 From governance over populations to governance with centaurs

The presence of ethical centaurs also changes the very style of governance that is possible. Instead of seeing AI fields as systems to be controlled from above, field-level institutions can treat centaurs as partners in identifying problems, experimenting with solutions, and monitoring impacts.

Governance over populations treats people primarily as objects of regulation and targets of policy. Rules are written, enforced, and adjusted based on aggregate indicators. Individuals and organizations comply or resist. Feedback is indirect and mediated. In such a regime, AI plays easily into existing patterns of centralized management; systems can be deployed to monitor compliance, nudge behavior, and optimize outcomes as defined by the governing body.

Governance with centaurs, in contrast, recognizes human–AI dyads as sites of situated expertise and moral judgment. It assumes that no central authority, however well-equipped, can fully anticipate the ways AI will reshape practices in diverse domains. Instead, it seeks to create frameworks within which centaurs can exercise discretion responsibly while contributing to collective learning.

Practically, this might involve regulatory experiments that are co-designed with centaur communities, iterative standards that evolve based on feedback from practitioners, or oversight bodies that draw heavily on reports from centaurs about emerging risks and unintended consequences. It may also involve devolving certain decisions to professional or local institutions that are closer to centaur practice, within broad field-level guardrails.

Such governance is slower and more dialogical than pure technocracy. It involves negotiation, compromise, and the recognition of plural values. But it is also more likely to produce alignment that is durable because it is rooted in how people actually live with AI. Ethical centaurs act as both beneficiaries and co-authors of field governance, rather than as mere subjects.

8.4 Criteria for when field-level intervention becomes appropriate

Insisting on “centaurs before fields” does not mean that field-level intervention must wait until some ideal state of centaur practice is achieved. There are real risks that demand coordination beyond the scale of individual dyads or organizations. The question is when and how to intervene in ways that reinforce, rather than undermine, the development of ethical centaurs.

Several criteria can help.

First, the presence of organized centaur communities. Before major field-level rules are set, there should be evidence that in key domains, centaurs exist, can communicate, and have some capacity to articulate their experiences. If such communities are entirely absent, priority may need to be placed on supporting their emergence—through training, infrastructure, and institutional space—rather than on tightly prescribing field-wide behaviors.

Second, the nature of the risk. Certain risks, such as catastrophic misuse or systemic vulnerabilities that cannot be addressed locally, may justify early field-level intervention even when centaur practice is immature. In such cases, however, interventions should be framed explicitly as provisional and should include mechanisms for incorporating centaur feedback as it becomes available. Blanket, inflexible rules set in the absence of centaur input are more likely to produce misalignment and resentment.

Third, the effect of intervention on centaur agency. Field-level measures should be evaluated not only on whether they reduce certain metrics of risk, but also on whether they support or erode ethical centaur practice. Does a proposed regulation incentivize thoughtful human oversight, or does it encourage box-ticking and deference to automated safeguards. Does a standard require organizations to document and empower human roles in AI-mediated decisions, or does it treat humans as interchangeable operators.

Fourth, the availability of alternatives. In some cases, problems attributed to “field misalignment” may be better addressed by strengthening centaur practice and local institutions rather than by imposing new field-level controls. For example, instead of mandating complex technical audits in all domains, it may be more effective to require and support professional bodies to develop and enforce centaur-focused guidelines specific to their fields.

Finally, the reversibility and adaptability of interventions. Given the novelty of AI fields, field-level governance should be designed with humility: policies should be revisable, standards should admit exceptions where justified, and institutions should have pathways to learn from centaur experience and adjust. Irreversible structural choices made before centaur practice has matured risk locking in patterns that are difficult to correct later.

Taken together, these criteria suggest that legitimate field governance will often be incremental and dialogical. It will move in step with the growth of ethical centaur populations, responding to risks that cannot be managed locally while leaving room for centaurs to experiment, refine their practices, and build the micro-institutions that make alignment real in everyday life.

The next sections turn from principles to more concrete design questions: how to cultivate more ethical centaurs, how different futures look depending on their density, and what kinds of objections and limits this centaur-first framing must confront.

9. Design Principles for Cultivating More Ethical Centaurs

If ethical centaurs are the primary unit of alignment, then the question becomes practical: how do we make them more common, more capable, and more resilient than their merely efficient or misaligned counterparts. This is not a matter of individual virtue alone. It depends on how tools are built, how people are trained, how organizations reward behavior, and what the research community chooses to optimize.

This section proposes four clusters of design principles. The first concerns interface and workflow designs that invite reflection instead of passive dependence. The second addresses training and education for centaur literacy and discipline. The third focuses on organizational incentives that recognize and reward accountable practice. The fourth turns to research priorities, arguing that centaur flourishing should become an explicit target for technical and social inquiry.

9.1 Interface and workflow designs that support reflection, not sedation

Interfaces are where centaur practice either comes alive or is quietly smothered. A system that always presents a single confident answer, with minimal context and no friction, nudges humans toward uncritical acceptance. A system that can present alternatives, show uncertainty, and make it easy to ask “why” supports a different posture.

Several design principles follow.

First, expose uncertainty and alternatives. When possible, systems should indicate degrees of confidence and offer multiple options rather than a single, apparently definitive output. This gives the human something to compare and reason about. Even a simple pattern—“Here are three possible approaches; here is where each might fail”—creates space for judgment.

Second, make it cheap to ask for explanation, and make explanations useful. An explanation button that produces canned, generic text does little for centaur practice. Explanations should be tied to the actual output and context: what data or patterns the system is leaning on, what

constraints it assumed, what trade-offs it implicitly made. Humans do not need full internal transparency; they need enough insight to decide when to trust, when to question, and what to check next.

Third, build “slow lanes” into workflows. High-throughput interfaces encourage users to move rapidly from one decision to the next. Ethical centaurs need ways to switch into slower, more deliberative modes when stakes are high. This might take the form of explicit flags that mark a case as sensitive, triggering additional prompts for review and documentation, or interface states that require extra confirmation steps.

Fourth, preserve manual paths. Some tasks should remain possible without AI, even if slower. Removing manual paths entirely forces users into reliance, making it harder to exercise veto power. Where full manual operation is impractical, systems can still provide modes where AI plays a reduced or advisory role, and humans take the lead.

Finally, avoid designs that blur responsibility. If an interface encourages mass acceptance of AI outputs with a single click, bundled into many other operations, users may lose track of which decisions they actually made. Interfaces that separate human approvals from automated actions, and that make those approvals visible in logs and timelines, reinforce the sense that the human remains accountable.

In sum, interface and workflow design can either sedate or stimulate the reflective capacities that ethical centaurs rely on. Designs that foreground speed and convenience above all risk producing efficient, but ethically hollow, human–AI configurations.

9.2 Training and education for centaur literacy and discipline

Most people are not taught how to think with AI. They are handed tools and told, implicitly or explicitly, to “be more productive.” If centaurs are to become ethical by default, training and education must change.

Centaur literacy starts with basic concepts: that AI systems are pattern learners, not oracles; that they can be systematically biased; that their behavior depends heavily on how tasks are framed; and that delegation decisions are themselves ethical choices. This literacy should be woven into professional education in fields where AI is likely to play a role, alongside domain-specific content.

Beyond concepts, there is centaur discipline: habits that sustain ethical practice under pressure. Training should include:

- Exercises in prompt design that highlight how subtle changes in framing affect outputs.

- Practice sessions in which participants deliberately probe system limits and failure modes.
- Scenarios that require deciding when not to use AI, even when it would be convenient.
- Role-playing exercises in explaining AI-assisted decisions to affected people.

Assessment should not focus only on technical proficiency but also on reflection. Can the practitioner articulate why they chose to accept one suggestion and reject another. Can they identify where they might be over-relying on the system. Can they describe the potential harms of delegating certain tasks.

For working professionals, continuing education is critical. As tools evolve, so will best practices. Organizations and professional bodies can offer workshops, guidelines, and peer discussion spaces focused specifically on centaur practice. Importantly, these should not be limited to technical staff. Managers, policymakers, and other decision-makers need centaur literacy as much as frontline practitioners.

Finally, education should also address the psychological dimension. Working closely with powerful AI can induce overconfidence, dependence, or alienation. A mature centaur curriculum acknowledges these possibilities and offers strategies for maintaining a healthy sense of agency and responsibility.

9.3 Organizational incentives that reward accountable centaur practice

Even the best interfaces and training will falter if organizational incentives pull in the opposite direction. If people are rewarded strictly for throughput, cost savings, or short-term metrics, they will have little reason to exercise the reflective, sometimes slower practices that ethical centaurs require.

To support centaur practice, organizations can adjust incentives in several ways.

First, recognize and reward responsible overrides. When humans choose to override or slow down an AI-mediated process because they perceive a risk or value conflict, this should not automatically be treated as inefficiency. Instead, organizations can track such events, investigate outcomes, and, when justified, commend the exercise of judgment. Over time, patterns in overrides can also guide system improvements.

Second, integrate centaur quality into performance evaluations. For roles that regularly interact with AI, evaluations can include criteria such as quality of documentation, thoughtful use of AI suggestions, engagement with training, and willingness to raise concerns about misalignment. These cannot be reduced to simple numbers, but they can be discussed in reviews and supported by supervisor observations and peer feedback.

Third, align management metrics with centaur goals. If managers are measured only on aggregate productivity and incident rates, they may unintentionally discourage the very practices that prevent subtle harms. Adding qualitative indicators—such as staff reports of feeling able to question systems, or periodic audits of centaur workflows—can help balance this.

Fourth, protect whistleblowers and critics. In any deployment, there will be pressures to downplay issues that cast AI use in a negative light. Organizations that are serious about ethical centaurs must ensure that those who raise legitimate concerns about AI-mediated practices are not punished or sidelined. Clear policies, independent reporting channels, and leadership signals matter.

Finally, allocate time and resources. Ethical centaurs take time: for reflection, for review, for learning. Organizations must make that time available. This may mean intentionally limiting the number of AI-mediated tasks assigned, restructuring workloads, or investing in support roles that help maintain centaur practices.

Without changes in incentives, centaur ethics will remain an individual burden. With them, it can become a shared organizational project.

9.4 Aligning research priorities with centaur flourishing

Research and development communities play a decisive role in shaping what kinds of human–AI relationships are even possible. If most effort goes into autonomy, throughput, and fine-grained control, centaurs will be an afterthought. If, instead, centaur flourishing becomes an explicit research priority, new tool designs and evaluation methodologies will emerge.

This alignment can take several forms.

First, develop metrics and benchmarks that capture centaur-relevant properties. Instead of focusing exclusively on model accuracy, latency, or benchmark scores, researchers can also measure how systems affect human decision quality, learning, trust, and error detection. Experiments can compare different interface designs, feedback mechanisms, and delegation patterns from the perspective of centaur outcomes.

Second, invest in tools that support centaur reflection. This includes research on interactive explanation, conversational debugging, and interfaces that make it easy to explore counterfactuals. It also includes approaches that allow humans to teach systems about their own values and constraints in interpretable ways, rather than only adjusting hidden parameters.

Third, study centaur practice in the wild. Empirical research on how people actually use AI systems in real settings—what shortcuts they take, where they feel pressured, how their skills change over time—can inform both design and policy. Such studies require collaboration across

disciplines: human–computer interaction, sociology, psychology, economics, and domain-specific expertise.

Fourth, support alternative infrastructures. If all AI use depends on a few centralized platforms, the diversity of centaur practice will be constrained. Research on open models, local deployment, and modular systems that smaller organizations can adapt supports a healthier ecology of centaurs with more varied configurations and sources of control.

Finally, reflect on research narratives themselves. The way the research community talks about AI—whether as replacement for human judgment, as a neutral tool, or as a partner—shapes expectations and funding. Emphasizing centaurs not only in technical papers but also in public communication can help normalize the idea that alignment is fundamentally about relationships, not just about systems.

In all these ways, research priorities can either nourish or neglect ethical centaurs. Making centaur flourishing a first-class goal does not preclude work on autonomy or efficiency, but it insists that these be pursued in a way that preserves and enhances human agency, rather than sidelining it.

The next section turns from design principles to longer-term trajectories, contrasting futures in which centaurs are rare and weak with those in which they are numerous and strong, and asking what these differences imply for the shape of AI fields and their governance.

10. From Centaurs to Fields: Conditional Futures

Ethical centaurs do not guarantee a good future, but their presence or absence shapes the kinds of futures that are even possible. Fields will form regardless: infrastructures will be built, regulations will be written, and economic incentives will push AI into every domain that can be automated or augmented. The open question is what these fields look like from the inside and whose agency counts.

This section sketches three conditional futures. Scenario A imagines monopolized fields with weak centaurs, where a small number of actors control the infrastructure and centaurs exist only as thin veneers over largely automated systems. Scenario B imagines federated fields grounded in strong centaur cultures, where human–AI dyads are central to practice and governance. Scenario C imagines fractured fields with isolated centaur enclaves, where alignment is achieved locally but fails to scale. The section closes by proposing that the density and health of centaurs should be treated as a leading indicator of which path a field is taking.

10.1 Scenario A: monopolized fields with weak centaurs

In Scenario A, AI fields are dominated by a small number of large providers who control critical infrastructure, model development, and distribution channels. Their systems are tightly integrated into major institutions: corporations, governments, and key service providers. AI is everywhere, but most human–AI relationships are shallow.

From the user's point of view, interaction with AI is mediated through uniform interfaces controlled by these providers. Tools are powerful and convenient, but opaque. Decisions about delegation boundaries, logging, and oversight are made centrally and encoded into platforms. Local practitioners have limited ability to customize workflows or challenge defaults.

Centaurs exist, but they are weak. A clinician or manager might technically have the option to override an AI recommendation, but time pressure, institutional policy, and interface design make such overrides rare. Training focuses on operational compliance rather than reflective practice. Those who push back against misuses or raise ethical concerns find little institutional support.

Field governance in this scenario is technocratic and corporatized. Regulations are negotiated between powerful firms and state actors, framed in terms of systemic risk, competitiveness, and security. Standards and audits emphasize model performance and incident rates, not centaur quality. Public participation is limited to consultation processes and occasional high-profile controversies.

This configuration can deliver short-term gains: increased efficiency, rapid deployment of new capabilities, and a degree of centralized control over obvious harms. But the alignment it offers is shallow. Misalignments at the level of human agency, local values, and institutional culture accumulate below the resolution of field-level metrics. Over time, people experience AI as something done to them rather than with them. Trust becomes fragile, and attempts to restore it through more control only deepen the sense of powerlessness.

Scenario A is attractive to those who favor decisive action and clear lines of authority. It is also the path of least resistance if ethical centaur practice remains underdeveloped. Without strong centaur cultures to anchor alternative possibilities, monopolized fields with weak centaurs are a plausible default.

10.2 Scenario B: federated fields grounded in strong centaur cultures

Scenario B imagines a different trajectory. Here, ethical centaurs are common in the roles and institutions that matter most, and their practices are recognized and supported. Infrastructures and governance arrangements are more federated: multiple providers, open standards, and diverse local configurations.

In this scenario, AI use in medicine, law, education, research, and public administration is typically organized as centaur practice. Professionals are trained not only in their domains but also in human–AI collaboration. They have meaningful control over how tools are integrated into their workflows, what tasks are delegated, and how decisions are documented and communicated.

Centaurs form dense networks and communities of practice. Their shared norms are reflected in organizational policies and professional guidelines. When field-level standards or regulations are proposed, centaur communities are actively consulted and have the capacity to respond with concrete feedback drawn from lived experience. They are not merely surveyed; they are co-authors of the rules.

Infrastructure in Scenario B is more plural. There may still be large providers, but there is also room for smaller platforms, open-source tools, and domain-specific systems governed by professional or civic institutions. Interoperability standards allow centaurs to move between tools or combine them in ways that fit local needs. No single provider has complete leverage over how AI is used in a domain.

Field governance in this scenario is dialogical and layered. Broad guardrails are set at higher levels to address catastrophic risks and cross-domain externalities, but much of the detailed governance is devolved to sectoral bodies, local institutions, and centaur communities. Policies are more likely to be iterative and revisable, with formal mechanisms for incorporating feedback from practice.

Alignment here is deeper but messier. Different sectors and regions may adopt different centaur norms, reflecting legitimate value pluralism. Conflicts and disagreements are common, but they are handled through ongoing negotiation rather than imposed settlement. Fields are less “neat” than in Scenario A, but they are more resilient: failures in one part of the system are less likely to cascade uncontrollably, and adaptation can occur at multiple scales.

Scenario B requires intentional cultivation. It depends on investments in training, institutional design, and infrastructure that support centaur agency, rather than simply scaling up centralized platforms. It also requires patience: federated fields grounded in centaur cultures evolve more slowly and may appear less efficient by narrow metrics. But they offer a better chance of aligning AI with the complexity of human values and institutions over the long term.

10.3 Scenario C: fractured fields and isolated centaur enclaves

Scenario C sketches a more ambivalent future. Here, strong ethical centaur cultures exist, but they are unevenly distributed and poorly connected. Some institutions, sectors, or regions manage to build healthy centaur practices and infrastructures. Others fall into patterns

resembling Scenario A, or into chaotic fragmentation where neither strong centaurs nor effective field governance exist.

In such a world, alignment becomes patchy. Within certain hospitals, universities, companies, or municipalities, AI is integrated in ways that respect human agency and support responsible practice. Centaurs in these enclaves have voice in local governance and may enjoy access to infrastructure that aligns with their values. Outside these islands, however, AI fields are monopolized, poorly regulated, or heavily captured by particular interests.

The result is a kind of alignment inequality. People's experience of AI and their ability to shape its use depend heavily on where they live, what institutions they belong to, and what economic resources they have. Ethical centaurs in well-resourced enclaves may be safe and effective, but they are limited in their ability to influence broader field trajectories. Attempts to export their norms or tools face resistance from entrenched interests or infrastructural incompatibilities.

At the field level, this scenario looks fractured. There is no coherent governance regime; instead, there are overlapping and sometimes conflicting rules imposed by different authorities. Some regions prioritize strict control, others laissez-faire innovation, and still others struggle to do anything coherent at all. Cross-border externalities proliferate: misaligned practices in one jurisdiction can harm people in another, and there is no agreed framework for resolving such disputes.

Centaurs in this scenario may feel both empowered and besieged. They know that better practice is possible, because they live it locally, but they see how difficult it is to scale or to protect against spillover from less aligned parts of the field. Some may withdraw into inward-focused optimization, giving up on broader influence. Others may become activists, trying to connect enclaves and build wider coalitions, but with limited bandwidth.

Scenario C is not purely dystopian; it preserves pockets of genuine alignment and innovation. But it is unstable. Strong centaur enclaves could, over time, network and move the field toward Scenario B, or they could be eroded or marginalized as Scenario A dynamics assert themselves. Much depends on whether centaur communities see themselves as part of a shared project across domains and borders, or as isolated experiments.

10.4 Measuring progress: density and health of centaurs as a leading indicator

Given these conditional futures, a key question for both practitioners and policymakers is how to tell which trajectory an AI field is on. Traditional indicators—such as model performance, market concentration, or the number of regulations passed—are insufficient. They say little about the underlying structure of human–AI relationships that will ultimately determine alignment.

This paper suggests that the density and health of ethical centaurs should be treated as a leading indicator. Density refers to how common centaur configurations are in roles and domains where AI has significant impact. Health refers to the quality of those configurations: whether humans retain real agency and responsibility, whether reflective practices are present, whether centaurs are supported by institutions and communities.

Measuring density might involve tracking, within a given sector:

- The proportion of key decisions made with AI support versus fully manually or fully automatically.
- The proportion of practitioners trained in centaur-specific skills.
- The presence of explicit human–AI collaboration policies in organizations.

Measuring health is more qualitative but still tractable. It might involve:

- Surveys of practitioners about their ability to question and override AI outputs.
- Audits of workflows to see whether time is allocated for reflection and review.
- Analysis of incident reports to determine whether centaur practices are catching and correcting errors.
- Studies of how centaur work affects skill development, job satisfaction, and perceived responsibility.

At the community level, indicators of centaur health could include:

- The existence and activity of centaur-focused communities of practice.
- The presence of shared prompts, protocols, and guidelines that embody ethical norms.
- The degree to which centaur perspectives are integrated into standards, audits, and policies.

Tracking these indicators over time can help diagnose whether a field is moving toward monopolized control with weak centaurs, toward federated governance grounded in strong centaur cultures, or toward fractured arrangements. They can also inform decisions about where to intervene: whether to invest in training, infrastructure, organizational reform, or field-level governance.

Most importantly, treating centaur density and health as a leading indicator keeps the focus where alignment is actually lived: in the everyday practices of people working with AI. It reminds designers, researchers, and policymakers that fields are not abstractions floating above

society; they are patterns emerging from countless human–AI dyads. To shape those patterns, attention must return, again and again, to the centaurs who inhabit them.

11. Objections, Limits, and Open Tensions

A centaur-first framing has intuitive appeal for those who value human agency and professional practice, but it is not without serious objections and limits. Some critics will see "ethical centaur" as little more than a flattering description of already privileged knowledge workers, offering a comforting story that leaves broader inequities largely intact. Others will point to relentless automation pressure and ask whether centaurs are economically sustainable outside a narrow band of jobs. Cultural and global differences further complicate any attempt to generalize centaur norms across societies. Finally, even in the best case, centaurs cannot by themselves resolve structural injustices or neutralize forms of hard power that operate well above the level of individual practice.

This section does not attempt to dissolve these tensions. Instead, it brings them to the surface, both to clarify the limits of the centaur concept and to prevent it from becoming an all-purpose solution in rhetoric while remaining marginal in reality.

11.1 Is "ethical centaur" just a romanticized knowledge worker

The first objection is that "ethical centaur" sounds suspiciously like a flattering rebrand of a particular type of worker: educated, desk-based, relatively autonomous, embedded in professional institutions. In this view, the centaur is simply a knowledge worker plus a new tool, described in elevated language that obscures the ordinary dynamics of labor and control.

There is some justice in this critique. Many of the examples used to illustrate centaur practice come from domains where individuals already enjoy significant status and discretion: doctors, lawyers, scientists, engineers. These are precisely the contexts in which there is space for reflective collaboration with AI, because roles are designed to include judgment and responsibility. It is easier to imagine a physician as an ethical centaur than a warehouse worker or delivery driver, whose tasks may be closely scripted and monitored.

If the centaur concept remains confined to these elite settings, it risks reinforcing existing inequalities. Those who already have agency and voice gain more powerful tools and richer vocabularies to describe their work, while others experience AI primarily as an instrument of surveillance, scheduling, or automated decision-making imposed from above. The centaur, in that case, becomes a romanticized figure in a narrow corridor of the economy, surrounded by large regions where human–AI relationships are instrumental and coerced.

To avoid this, two moves are necessary. First, the centaur framing must be applied critically to contexts where autonomy is constrained. It should ask not only “How can we help privileged professionals think better with AI” but also “Where could meaningful centaur practice exist in roles currently designed for minimal discretion, and what would have to change.” That may involve redesigning jobs, not just tools.

Second, the centaur should not be treated as a status marker but as a normative demand: wherever AI is used to mediate consequential decisions about people’s lives, there should be a human–AI configuration in which some person, or group, genuinely inhabits the role of ethical centaur. If that is impossible in a given workflow, the question becomes whether that workflow should exist at all in its current form. In this way, the centaur concept can serve as a test for the legitimacy of AI-mediated arrangements, not a celebration of knowledge workers alone.

11.2 Automation pressure and the economics of replacing centaurs

A second objection is more blunt: centaurs may be ethically appealing, but they are economically fragile. If a fully automated system can perform a task more cheaply and quickly than a centaur, organizations under competitive pressure will be tempted to remove the human, or to reduce their role to a minimal rubber stamp. Over time, centaurs could be squeezed out except in niches where regulation or liability makes human involvement mandatory.

Automation pressure is real. In many domains, the short-term financial incentives point toward maximizing throughput and minimizing labor costs. Ethical centaurs are, by design, slower and more expensive than pure automation in some cases, because they allocate time for reflection, review, and documentation. From a narrow cost perspective, this can look like inefficiency.

The centaur-first argument responds in two ways. First, it insists that some tasks are not legitimate candidates for full automation, regardless of apparent cost savings. Decisions that affect fundamental rights, bodily integrity, or long-term life chances require a human locus of responsibility. In these areas, centaurs should be treated not as optional upgrades but as minimum conditions for justifiable AI use. This is ultimately a political and legal claim, not an economic one.

Second, it points out that automation pressure often underestimates long-term costs: erosion of trust, skill decay, vulnerability to systemic failures, and the difficulty of correcting course once fully automated pipelines are entrenched. Centaurs can be seen as investments in resilience and corrigibility. Proving this requires work: empirical studies that compare centaur and fully automated regimes over time, including indirect costs and externalities. Without such evidence, appeals to resilience will sound vague next to immediate cost savings.

Even with these arguments, there will be domains where centaurs are displaced. The centaur-first perspective does not promise to halt automation everywhere. It seeks to draw lines around where centaur practice is non-negotiable and to encourage organizations and regulators to value long-term stability and legitimacy alongside efficiency. Some tension with short-term economics is unavoidable.

11.3 Cultural and global variation in centaur norms

A third objection concerns universality. The centaur concept, as developed here, reflects assumptions about individual agency, professional roles, and governance that are not globally shared. Expectations about deference to authority, collective versus individual responsibility, and the proper role of technology vary widely across cultures and political systems. What counts as an "ethical centaur" in one context may look irresponsible or alien in another.

For example, in some settings, it may be considered appropriate for AI systems to embody and enforce state-sanctioned norms with minimal room for individual dissent. In others, strong hierarchical institutions may view centaur-like practices—where individuals question or override system recommendations—as threats to order. Conversely, in contexts with histories of technological harm or political oppression, skepticism of AI may be so strong that centaur practice is seen as complicit with an untrusted infrastructure.

These variations mean that centaur norms cannot simply be exported wholesale as a global template. They must be adapted to local institutions, histories, and value frameworks. The core elements of the centaur—shared practice, human veto, reflective use—may still be relevant, but their expression will differ. In some places, centaurs may be anchored in community institutions rather than individual professionals. In others, religious or philosophical traditions may shape what counts as legitimate delegation to machines.

At the same time, the use of AI in transnational infrastructures—cloud platforms, global supply chains, cross-border financial systems—creates pressures for some degree of convergence. When a decision made by a centaur in one jurisdiction affects people in another, disagreements about acceptable practice become more than academic. Negotiating these tensions will require new forms of international dialogue, in which centaur communities from different cultures participate alongside states and firms.

The centaur-first framing does not resolve global pluralism. It suggests that any serious attempt to govern AI fields must contend not only with differences in law and policy but with differences in how people conceptualize human–machine collaboration and moral responsibility. Ethical centaurs will look and feel different in different parts of the world; taking that seriously is part of the work.

11.4 What centaurs cannot solve: structural injustices and hard power

Finally, there is a set of limits beyond which centaurs simply do not reach. Human–AI dyads, however well designed, cannot by themselves undo structural injustices or neutralize hard power. They operate within economic systems, political regimes, and cultural orders that predate AI and that shape both the goals and the constraints of centaur practice.

A clinician functioning as an ethical centaur can offer more thoughtful, person-centered care within a health system that remains deeply unequal. A manager practicing centaur ethics can soften the impact of certain policies while still operating in an organization whose business model depends on precarious labor. A public servant centaur can resist abuses at the margins while still working inside a state apparatus capable of coercive force.

Centaurs can push back, signal problems, and sometimes refuse. They can form networks that advocate for changes in policy or infrastructure. But they cannot, on their own, redistribute wealth, alter geopolitical power balances, or rewrite constitutional structures. If those structures are fundamentally misaligned with widely held values, centaur practice may mitigate harms, but it will not resolve the underlying tensions.

There is also the blunt reality of coercive power. In contexts where dissent is criminalized, where surveillance is pervasive, or where violence is used to enforce conformity, ethical centaurs may be impossible or dangerous. AI systems can be deployed explicitly as instruments of control, and those who refuse to collaborate with them may face serious consequences. In such situations, centaur language may be co-opted as a veneer over coercive practices, or it may simply be absent.

Recognizing these limits is important to avoid overburdening the centaur concept. Ethical centaurs are one layer in a multi-layered alignment problem. They matter because they are where many everyday decisions are made and where lived experience of AI is shaped. But they cannot substitute for broader struggles over justice, democracy, and the distribution of power.

A centaur-first approach to alignment, therefore, should be seen as a complement to, not a replacement for, work at other levels: institutional reform, legal and regulatory change, and political movements. It asks that wherever AI is integrated into human activity, attention be paid to the quality of the human–AI relationships that result. It does not claim that this alone will save us.

In the conclusion that follows, the paper returns to its central thesis: that sequencing matters, that centaurs must come before fields, and that any attempt to govern AI at scale must remain anchored in the everyday moral labor of human–AI dyads.

12. Conclusion: First People, Then Fields

This paper has argued for a reordering of how we think about AI alignment and governance. Rather than treating the progression from models to stacks to fields as a simple expansion in scale and abstraction, it has insisted on inserting another unit into the story: the ethical centaur, the human–AI dyad in which responsibility, judgment, and power are actually lived. The central claim is simple to state and difficult to honor in practice: fields should not be designed or governed over people whose everyday relationships with AI are shallow, coerced, or unformed. The people–machine partnerships must come first.

Centaurs are not a sentimental add-on to field-level analysis; they are the substrate from which any legitimate field must grow. Without them, field governance can at best optimize visible metrics and manage large-scale risks, while leaving unaddressed the subtler ways in which AI reshapes agency, norms, and institutions. With them, fields can be more than managed infrastructures; they can be environments in which human intelligence and moral imagination are genuinely amplified rather than sidelined.

The conclusion of a conceptual paper cannot offer final answers. Instead, it can restate the ordering, make explicit the way ethical centaurs act as living checks on technocratic drift, suggest concrete directions for research and institutional design, and end with a caution and an invitation to those now in a position to shape AI futures. The ordering is normative, not inevitable. Whether it is realized depends on choices made now by designers, practitioners, organizations, and policymakers.

12.1 Reasserting the ordering: why centaurs must come first

“Centaurs before fields” is, at its core, a claim about where alignment work actually lives. Models and stacks are necessary technical substrates. Fields are emergent environments in which those substrates interact with infrastructures, markets, and regulations. But the place where values are interpreted, where delegation decisions are made, where harm is perceived, and where responsibility is felt is the human–AI dyad.

If we skip over this dyad and go directly from stacks to fields, we trade lived complexity for administrative clarity. We gain dashboards, regulations, and strategies at the cost of losing sight of how people actually inhabit AI-saturated systems. The result may be impressive in policy documents and corporate roadmaps, but brittle in practice, because it rests on thin assumptions about human behavior and institutional adaptation.

Putting centaurs first does not mean freezing fields until a perfect centaur culture emerges. It means refusing to treat field-level governance as the primary site of alignment until there is at least a recognizable population of ethical centaurs in key domains, with some training,

institutional support, and channels for voice. It means designing early governance measures in ways that nurture centaur agency rather than crowding it out.

This ordering also serves as a diagnostic. In any proposal for AI regulation, infrastructure investment, or large-scale deployment, one can ask: where are the centaurs in this picture. How are they trained. What authority do they have. How will their experiences and objections reach the level where rules are made. If the answers are vague or absent, the proposal is misaligned with the ordering this paper defends, regardless of how sophisticated it looks at the field level.

12.2 Ethical centaurs as living checks on technocratic drift

Technocratic drift occurs when decisions about complex systems are made ever more exclusively by those who control models, infrastructures, and formal governance processes, using tools and metrics that make sense from their vantage point but may poorly reflect the lives of those subject to the systems. AI fields are particularly vulnerable to this drift because of their opacity, speed, and scale.

Ethical centaurs are one of the few available counterweights. They experience, at close range, how AI tools interact with real situations, values, and institutional constraints. They see where systems help and where they distort practice. They are positioned to notice misalignments long before they surface in aggregate statistics or public crises. They can refuse, modify, or reinterpret AI outputs in ways that preserve human dignity and responsibility.

For centaurs to function as checks on technocratic drift, their role must be more than individual conscience. They need communities of practice where their experiences are shared and analyzed, organizational structures that protect their ability to say no, and governance frameworks that invite their input rather than treating them as mere implementers. In other words, centaur practice must be recognized as epistemically valuable, not only operationally necessary.

When such recognition exists, ethical centaurs can act as living sensors and governors embedded throughout the field. They can respond locally and quickly to emerging misalignments, and they can feed those insights upward into institutions capable of adjusting broader rules. They are imperfect and partial, but they are responsive in a way that top-down field management rarely is. Ignoring them—or allowing their role to be eroded by automation and rigid compliance regimes—removes one of the few mechanisms available to keep AI fields from sliding into purely managerial forms of control.

12.3 Directions for research, practice, and institutional innovation

Taking “centaurs before fields” seriously generates a wide agenda for research, practice, and institutional innovation. It suggests that alignment work should be measured as much in the

quality and prevalence of ethical centaurs as in the safety properties of models or the sophistication of regulations.

For research, this means studying centaurs directly: how different interface designs shape their judgment, how their skills evolve over time, how they perceive responsibility and risk, and how their practices aggregate into proto-fields. It means developing evaluation frameworks that treat human–AI dyads as the unit of analysis, rather than only isolated models or entire organizations. It calls for interdisciplinary collaboration between technical researchers, social scientists, ethicists, and practitioners in specific domains.

For practice, it means investing in centaur literacy and discipline. Professional education and workplace training should incorporate not only tool operation but also norms of delegation, veto, explanation, and withdrawal. Organizations should treat centaur practice as a core competency, subject to continuous improvement, not as an informal layer on top of existing workflows.

Institutional innovation is perhaps the hardest piece. New or adapted bodies may be needed to represent centaur interests in standards and policy processes. Professional associations could develop centaur-specific codes of practice. Regulators might experiment with oversight mechanisms that rely on centaur reporting and review, rather than solely on technical audits or compliance checklists. Funding structures could prioritize projects that explicitly strengthen centaur agency and networks.

None of these directions replaces work on model robustness, interpretability, or field-level governance. They complement it, ensuring that technical advances and regulatory frameworks are anchored in the everyday realities of human–AI collaboration.

12.4 A caution and an invitation for those who would shape AI futures

The caution is straightforward: it is entirely possible to build impressive AI fields with only weak, coerced, or absent centaurs. The economic and political incentives to do so are strong. Centralized platforms, automated decision pipelines, and technocratic governance architectures are easier to design, measure, and sell than messy, federated arrangements grounded in diverse human–AI dyads. A future in which AI is everywhere and centaur agency is rare is plausible.

The invitation is equally clear: those now in positions to shape AI—researchers, engineers, entrepreneurs, professionals, regulators, educators—can choose otherwise. They can design models, stacks, and institutions with centaur flourishing as an explicit goal. They can insist that major deployments include plans for training and supporting ethical centaurs, not only for deploying infrastructure. They can use their influence in standard-setting and policy to create space for centaur voices, not just for technical and corporate perspectives.

This is not a heroic appeal to individuals to single-handedly resist large systems. It is a call to orient collective work toward the human–AI relationships in which alignment will actually be lived. If, in a decade, we look back on the early years of widely deployed AI and find that we built fields first and only later wondered why people felt disempowered and misaligned, the mistake will not have been a lack of foresight. It will have been a failure to prioritize the ordering that was already visible: first people, then fields.

Centaurs are not a guarantee of good outcomes, but they are a necessary condition for any AI future that stays recognizably human in its distribution of agency and responsibility. To put them first is to acknowledge that alignment is not only a question of engineering and policy, but of how we choose to inhabit our tools—and how we allow our tools to inhabit us.

13. References

The following references are grouped by theme. They are not exhaustive, but they give readers a starting map across human–AI collaboration, socio-technical systems, governance, and systemic risk.

13.1 Work on human–AI collaboration, centaurs, and hybrid intelligence

Kasparov, G. (2017). Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins. PublicAffairs.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. Business & Information Systems Engineering, 61(5), 637–643.

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2019). The Future of Human–AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. In Proceedings of the 52nd Hawaii International Conference on System Sciences.

Saghafian, S., & Idan, L. (2023). Effective Generative AI: The Human-Algorithm Centaur. Harvard Kennedy School Faculty Research Working Paper Series, RWP23–030.

Puranam, P. (2023). Humans and AI in Organizations: Hybrid Intelligence in Practice. In A. Foss & H. J. Larson (Eds.), The Future of Management in an AI World.

These works collectively motivate the centaur and hybrid-intelligence lens by examining how humans and AI systems jointly contribute to cognition, decision-making, and organizational performance. Harvard Kennedy School+6aitskadapa.ac.in+6The Guardian+6

13.2 Literature on socio-technical systems and large technical infrastructures

Hughes, T. P. (1983). Networks of Power: Electrification in Western Society, 1880–1930. Johns Hopkins University Press.

Mayntz, R., & Hughes, T. P. (Eds.). (1988). The Development of Large Technical Systems. Westview Press.

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. Information Systems Research, 7(1), 111–134.

Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. Social Studies of Science, 14(3), 399–441.

Trist, E. (1981). The Evolution of Socio-Technical Systems: A Conceptual Framework and Action Research Program. Occasional Paper No. 2, Ontario Quality of Working Life Centre.

Winner, L. (1980). Do Artifacts Have Politics? Daedalus, 109(1), 121–136.

This body of work develops the conceptual tools for thinking about AI not only as software or hardware, but as part of wide socio-technical fields and infrastructures. sistemas-humano-computacionais.wdfiles.com+6Hopkins Press+6Econstor+6

13.3 Governance, regulation, and ethics of AI and adjacent technologies

Floridi, L. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1). Harvard Data Science Review

Floridi, L. (2023). The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford University Press. OUP Academic

High-Level Expert Group on Artificial Intelligence. (2019). Ethics Guidelines for Trustworthy AI. European Commission. European Parliament+1

Corrêa, N. K., et al. (2023). Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance. Patterns, 4(11), 1–23. ResearchGate+3ScienceDirect+3arXiv+3

Taeihagh, A. (2025). Governance of Generative AI. Policy and Society, 44(1), 1–22. OUP Academic+2ResearchGate+2

Dicastery for Communication & Dicastery for Culture and Education. (2025). Antica et nova: On Artificial Intelligence and Human Dignity. Vatican City. Reuters

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. UNESCO General Conference, 41st Session. OUP Academic

OECD. (2019). OECD Principles on Artificial Intelligence. Organisation for Economic Co-operation and Development. ScienceDirect

These references trace the emerging landscape of AI principles, regulatory proposals, and institutional guidelines that will shape the conditions under which centaurs and fields develop.

13.4 Empirical and theoretical studies on alignment, agency, and systemic risk

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press. WIRED+4Amazon+4Oxford University Press+4

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking. Wikipedia+1

Yudkowsky, E. (2016). The AI Alignment Problem: Why It's Hard, and Where to Start. Talk at the 26th Annual Symbolic Systems Distinguished Speaker Series, Stanford University; transcript published by the Machine Intelligence Research Institute. Stanford Humanities Center+3MIRI+3MIRI+3

Nielsen, M. (2025). ASI Existential Risk: Reconsidering Alignment as a Goal. Astera Institute. Michael's Notebook+1

Ord, T. (2020). The Precipice: Existential Risk and the Future of Humanity. Bloomsbury. Wikipedia+1

Perrow, C. (1999). Normal Accidents: Living with High-Risk Technologies (Updated Edition). Princeton University Press. JSTOR+2Internet Archive+2

Beck, U. (1992). Risk Society: Towards a New Modernity. Sage. Internet Archive+2WorldCat+2

Schwerzmann, K., & colleagues. (2025). Desired Behaviors: Alignment and the Emergence of a New Class of AI Risks. Journal of Artificial Intelligence and Ethics, 5(2), 101–130. PMC

Collectively, these works connect AI alignment and agency to broader theories of systemic and existential risk, offering frameworks for thinking about how misalignment, power, and complex infrastructures interact at scale.

The author has published over 4,600 AI-assisted papers at: https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.