

First People, Then Centaurs: Joint Human–AI Agency as the Next Evolutionary Step

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

December 8, 2025

The rapid ascent of large language models has made it tempting to speak as if the next step in evolution belongs to machines alone. Yet the more consequential transformation may lie not in disembodied models, but in a new kind of joint creature: the centaur. A centaur is not simply a “power user” of AI tools. It is a stable union between a human being who retains moral responsibility and a layered AI stack that expands perception, memory, and reasoning without erasing conscience or skill. This paper argues that exploration of the centaur must become a central theme before we attempt any serious move toward AI “fields” – pervasive, infrastructural systems that shape everyday life. Moving too quickly from models to fields risks solidifying patterns where humans become pets of opaque systems or subjects of AI-augmented oligarchs. By contrast, centaur practice foregrounds joint agency, reversibility, and non-disposability of persons. We develop a conceptual ladder from model to stack to centaur to field, identify the pathologies that emerge if the centaur step is skipped, and propose ethical and institutional conditions under which centaur patterns can flourish. The goal is not to glorify technology, but to preserve human dignity as intelligence scales.

Keywords: centaur intelligence, human AI collaboration, AI stacks, field-level AI, joint agency, infrastructural love, ethical deployment, non disposability, alignment, hybrid cognition, technological evolution

A collaboration with GPT-5.1-Thinking-Extended. 47 pages. CC4.0.

Outline

1. Introduction: From Models to Joint Creatures

- 1.1 The current narrative of AI evolution
- 1.2 Reframing progress around relationships, not just capabilities
- 1.3 Overview of the model → stack → centaur → field ladder
- 1.4 Scope and aims of the paper

2. Four Levels of Deployment: Model, Stack, Centaur, Field

- 2.1 Models as general-purpose engines of transformation
- 2.2 Stacks as governed systems around models
- 2.3 Centaurs as unions of humans and stacks with shared work
- 2.4 Fields as ambient, infrastructural AI environments
- 2.5 Why the arrows between levels are not automatic

3. Defining the Centaur: Beyond Power Users and Tools

- 3.1 Centaur as a role, not a metaphor
- 3.2 Distinguishing centaurs from tool users, pets, and operators
- 3.3 Joint agency and the retention of moral responsibility
- 3.4 The inner life of a centaur: discernment, doubt, and conscience

4. Evolutionary Forks: Pets, Oligarchs, or Centaurs

- 4.1 The pet pattern: humans as front-ends for hidden decisions
- 4.2 The oligarch pattern: AI as an exoskeleton for concentrated power
- 4.3 The centaur pattern: shared power with preserved human depth
- 4.4 Why centaur exploration is the only humane next step
- 4.5 Skipping the centaur: structural consequences for future fields

5. Empirical Centaurs: Where Joint Agency Already Exists

- 5.1 Emerging patterns in writing, research, and knowledge work
- 5.2 Clinical, educational, and creative centaurs
- 5.3 Strong versus weak centaur practices
- 5.4 Lessons from early adopters and implicit centaur cultures

6. Norms and Oaths: What a Good Centaur Refuses

- 6.1 Offloading conscience versus using AI to see more clearly
- 6.2 Refusing to hide behind “the model said”
- 6.3 Designing against deskilling and moral atrophy
- 6.4 A draft centaur oath: commitments and prohibitions
- 6.5 The role of formation, not just training

7. Infrastructural Love: Constraints on Centaur and Field

- 7.1 Love at scale as non-disposability and care in structure
- 7.2 The “Beautiful” constraint as a deployment test
- 7.3 Interfaces that reveal structure, uncertainty, and limits
- 7.4 Reversibility, kill-switches, and the right to say no
- 7.5 Embedding these constraints in law, design, and culture

8. Preparing Institutions for Centaur Cultures

- 8.1 Rethinking education: training centaurs, not just users
- 8.2 Professional roles and governance for centaur practice
- 8.3 Incentives that reward judgment over blind efficiency
- 8.4 Documentation, audit, and shared responsibility in centaur teams
- 8.5 Institutional examples and sketches for implementation

9. From Centaurs to Fields: Conditions for Ethical Expansion

- 9.1 When centaur patterns are mature enough to scale
- 9.2 Avoiding monocultures and brittle dependencies
- 9.3 Civic legitimacy, opt-out paths, and public debate
- 9.4 Designing fields that bend around centaur dignity
- 9.5 Failure scenarios and criteria for rolling back field experiments

10. Conclusion: First People, Then Centaurs

- 10.1 Summary of the argument for centaur centrality
- 10.2 What is at stake if centaur exploration is neglected
- 10.3 Directions for research, practice, and public discourse
- 10.4 A modest ambition: instruments that do not trample the people they serve

11. References

1. Introduction: From Models to Joint Creatures

The story most people hear about artificial intelligence is a story of ascent. Each new generation of systems is framed as a higher rung on a ladder: bigger models, more parameters, better benchmarks, more “general” capabilities. In this narrative, humans appear mostly as spectators or at best as temporary teachers, feeding data into an increasingly autonomous intelligence that will eventually march off on its own. Progress is measured in leaderboards, not in the quality of relationships between humans and machines.

This paper takes a different starting point. Instead of treating AI as a standalone entity slowly replacing human minds, it treats the central object of interest as a hybrid: a human being working in sustained partnership with an AI system. This union, often called a centaur, is more than a user with a powerful tool. It is a joint creature that develops its own habits, strengths, and vulnerabilities. The centaur is where the most important questions of responsibility, dignity, and love will actually play out, long before any abstract “superintelligence” arrives.

To make this concrete, we will describe a ladder of deployment that runs from models to stacks to centaurs to fields, and then argue that exploration of the centaur step must become central if we care about humane evolution. The question is no longer only what these systems can do, but what kinds of beings we are becoming when we work with them.

1.1 The current narrative of AI evolution

In the dominant narrative, AI evolves along a relatively simple axis. Early systems are narrow and brittle, good at one task and useless outside it. Newer models are broader and more flexible, able to handle language, images, code, and control tasks under a single roof. The story is usually told in terms of:

- Scale: larger datasets, more parameters, more compute
- Performance: higher scores on benchmarks, better tool use, fewer obvious mistakes
- Automation: more tasks that no longer require human intervention

Within this frame, the natural endpoint is obvious. As models improve, they supposedly move toward full autonomy. Humans, in turn, become supervisors, safety checks, or legacy users for a shrinking set of activities. The real action, in this view, lies inside the model weights and training runs. Human lives are downstream beneficiaries or collateral.

There are at least three problems with this picture. First, it treats “intelligence” as a single substance that can migrate from biological to digital form without changing the nature of social life. Second, it hides the fact that every model is, in practice, deployed inside a human-led organization with particular incentives, blind spots, and power structures. Third, it encourages a

passive attitude among users: if the model is where the evolution happens, ordinary people can do little but adapt.

This paper starts from the observation that, in everyday practice, most powerful AI systems are not free agents. They exist as components in larger arrangements: chat interfaces, integrated tools, workstations, workflows, and institutional structures. The lived reality of AI today is not a model floating in isolation, but a composite of software, hardware, business logic, policy layers, and human users. That is where evolution is actually happening.

1.2 Reframing progress around relationships, not just capabilities

If we shift attention from isolated systems to the relationships they enter, a different picture emerges. The central variable is no longer only “how smart is the model,” but “what kind of relationship does this setup invite between human and machine.” Does it encourage deference or discernment, dependency or growth, disposability or care?

Progress then becomes multi-dimensional. A new generation of AI technology is not an upgrade simply because it can answer more questions or write longer code. It is an upgrade only if it makes possible better relationships:

- Relationships in which humans remain morally responsible for decisions, not shielded by machine authority
- Relationships in which human skills are deepened and extended, not quietly replaced and allowed to atrophy
- Relationships in which people can say no, change course, and step away without losing their ability to act in the world

In this reframing, the union between a human and an AI system is not a temporary convenience; it is a new kind of creature. The human brings history, conscience, embodiment, vulnerability, and the capacity to suffer and to love. The AI system brings speed, recall, pattern sensitivity, and tireless availability. Together they form a centaur whose behavior cannot be reduced to either component alone.

The quality of this centaur depends not only on technical performance but on design choices. Interfaces can be built to expose reasoning and uncertainty, inviting human judgment, or they can be built to present outputs as oracular, inviting passive acceptance. Organizational cultures can reward thoughtful oversight, or they can reward throughput and obedience to dashboard metrics. The same model can, under different conditions, support very different kinds of human growth or decay.

By placing relationships at the center, we make room for questions that are often sidelined. What happens to a person's sense of purpose as they lean on a stack every day? How does daily cooperation with an always-present system reshape character, attention, and empathy? How do these hybrid beings treat others who do not have similar access? These are evolutionary questions, even though they are asked in the language of ethics and culture rather than biology.

1.3 Overview of the model → stack → centaur → field ladder

To talk clearly about these dynamics, it helps to distinguish four levels at which AI participates in social reality:

- Model: a trained system that transforms inputs into outputs according to learned patterns
- Stack: a model embedded in tools, interfaces, data pipelines, policy layers, monitoring, and control mechanisms
- Centaur: a durable, patterned partnership between a particular human and a particular stack, with shared work and mutual adaptation
- Field: a pervasive, infrastructural presence of many stacks and centaurs woven into everyday environments, institutions, and flows of life

The arrow from model to stack represents a shift from raw capability to governed system. A model becomes a stack when it is wrapped in the instrumentation, constraints, and procedures that make it usable in the world: logging, rate limiting, safety filters, integration with other tools, and so on.

The arrow from stack to centaur represents a shift from system to relationship. A stack becomes part of a centaur when a human being uses it consistently enough that joint habits emerge. Over time, the human learns when to trust and when to question the system; the system may be fine-tuned or configured around that person's preferences and needs. The two become, for certain tasks, a single effective agent.

The arrow from centaur to field represents a shift from individual pattern to ambient infrastructure. A field arises when these hybrid arrangements are no longer rare or optional but become a background condition. Decision-making in organizations, education, healthcare, governance, and daily life is then shaped continuously by the presence of centaurs and stacks, often in ways that are difficult to see.

This ladder is not a simple progression of technical power. Each step introduces new risks and responsibilities. A model can be dangerous inside a badly designed stack. A stack can be harmful when it encourages weak or unethical centaur patterns. A field can become oppressive if it is

structured in ways that erode human agency and dignity. The main argument of this paper is that the centaur step cannot be skipped if we want any future field to be humane. Exploring, naming, and cultivating centaur patterns must come before attempts to saturate environments with AI.

1.4 Scope and aims of the paper

The scope of this paper is conceptual and normative rather than mathematical or empirical, although it draws on practical experience with emerging uses of AI in knowledge work and other domains. The aim is to offer a vocabulary and a set of distinctions that help orient decisions about development and deployment.

Specifically, the paper pursues four main goals.

First, it clarifies the four levels of model, stack, centaur, and field, emphasizing that these are not mere marketing labels but structurally different phases of deployment. Each level brings distinct kinds of risk and opportunity that cannot be reduced to “better models.”

Second, it develops a robust definition of the centaur as a role and a joint creature, distinguishing it from other patterns such as tool use, pet-like dependency, and oligarchic control. This includes an account of what it means for a human to retain moral responsibility while working closely with a powerful stack.

Third, it argues that exploration of the centaur must become a central theme in research, policy, and culture if we hope to avoid futures in which humans are either domesticated by opaque systems or outmatched by a small class of AI-augmented elites. This argument is grounded in an ethical stance that treats human beings as non-disposable and insists that love, care, and truthfulness must inform the architecture of our systems.

Fourth, it sketches conditions under which centaur patterns can mature into something worth scaling, and outlines safeguards for any subsequent move toward field-level deployment. These conditions include reversibility, transparency, civic legitimacy, and what might be called infrastructural love: the embedding of non-disposability and care into the very structure of technological and institutional life.

The paper does not attempt to predict specific timelines for AI development, nor does it offer a single blueprint for institutions. Instead, it proposes that the most important evolutionary step now available is not a leap in machine intelligence alone, but the cultivation of a new kind of joint human–AI being. The rest of the text explores what that means and what it would take to pursue it responsibly.

2. Four Levels of Deployment: Model, Stack, Centaur, Field

To understand where joint human–AI creatures come from, it helps to separate what is often blurred together. The same underlying model can exist in very different states of social reality. As a research artifact, it is little more than a complex function approximator. As part of a product, it becomes entangled with tools, business rules, and monitoring. In sustained cooperation with a person, it becomes a component of a hybrid agent. At scale, thousands of such hybrids and systems can shape whole institutions and environments.

This section clarifies the four levels of deployment that will recur throughout the paper: model, stack, centaur, and field. Each level is more than a technical phase; it is a different way of locating AI within human life. The harms and benefits that matter most are not symmetric across these levels, and the ethical questions change as one moves upward. Crucially, it is not inevitable that every model becomes a stack, every stack becomes a centaur, or every centaur world becomes a field. Those transitions are choices, not fate.

2.1 Models as general-purpose engines of transformation

At the base level, a model is a general-purpose engine for transforming inputs into outputs. It encodes statistical regularities learned from training data into a set of parameters. When prompted, it produces continuations, classifications, actions, or control signals that reflect those learned patterns.

In this state, the model has three key characteristics. First, it is context-poor with respect to the world in which it will eventually be used. It has seen many examples but knows nothing about which future uses are appropriate or dangerous. Second, it is largely uninterpreted from a human perspective. Engineers may have a sense of its capabilities and failure modes, but its internal representations are opaque even to its creators. Third, it is disembodied from governance. There are no built-in mechanisms for logging, rate limiting, or enforcing organizational policy. It is like a powerful mathematical object sitting on a lab bench.

As long as a model remains in this state, its impact is limited. Researchers can probe, test, and characterize it, but it has no direct interaction with users or institutions. This stage is important for safety work and capability evaluation, yet it does not by itself determine how the model will change lives. That depends on how it is embedded and constrained in concrete systems. To see that, one has to move up a level.

2.2 Stacks as governed systems around models

A stack is what emerges when a model is wrapped in everything necessary for real-world operation. This includes interfaces that humans interact with, tools the model can call, data pipelines and storage, monitoring and analytics, safety filters, and business logic. If the model is

a complex function, the stack is an engineered environment that decides when and how that function is invoked.

Stacks introduce several important forms of governance. They define who can access the model and under what conditions. They record queries and outputs, enabling audits and incident response. They implement guardrails to block certain categories of content and behavior. They provide mechanisms for updating, rolling back, or replacing the underlying model. In short, a stack moves from raw capability to managed service.

This governance is never neutral. The structure of a stack reflects the incentives and values of its owners. A stack built primarily for advertising will track attention and optimize for engagement. A stack built for clinical support will prioritize reliability, documentation, and integration with existing workflows. Two stacks built on the same underlying model can therefore have very different social consequences.

It is at the stack level that many of the public debates about safety and regulation arise. Questions about data retention, content moderation, consent, and liability attach to the system as a whole, not to the abstract model. Nonetheless, the stack is still, in essence, a tool. It responds to inputs according to design, but it does not yet define a particular way of being in the world for any specific human. That requires repeated interaction over time, which brings us to centaurs.

2.3 Centaurs as unions of humans and stacks with shared work

A centaur emerges when a particular human being works with a particular stack so consistently that they form a joint pattern of action. The components are familiar, but the result is new. The human brings goals, values, situational awareness, and embodied experience. The stack brings speed, memory, and pattern recognition. Over time, the two adapt to one another.

This adaptation has several dimensions. The human learns which kinds of questions the stack answers well, and which require extra scrutiny. They develop habits of prompting, checking, and integrating outputs into their own thought processes. The stack, in turn, may be tuned or configured to better match the person's domain, style, and preferences. Plugins and tools are added that map to the human's work. A shared history accumulates in the form of drafts, projects, and routines.

The result is a union with shared work. For many tasks, it no longer makes sense to ask whether it was "the human" or "the AI" who did the job. What matters is how the centaur as a whole behaved. Did it treat others fairly? Did it respect confidentiality? Did it show appropriate caution where stakes were high? Did it notice when the underlying model failed in a subtle way and compensate?

In this framing, the centaur is where moral responsibility resides. The human partner remains ultimately answerable for the actions taken. They cannot hide behind the stack as an excuse, because it is the human who chooses how to use it, when to override it, and whether to walk away. At the same time, it would be dishonest to ignore the fact that the stack shapes what the human notices, considers, and deems possible. The union is a genuine joint creature, sharing both power and vulnerability.

Centaurs already exist in practice. Writers who draft with models, clinicians who consult decision support tools, teachers who co-design lessons with AI assistants, and researchers who explore hypotheses with simulation and language models all rely on hybrid agency. Some of these centaur patterns are careful and life-giving; others are careless and dangerous. Naming the centaur helps bring these differences into focus.

2.4 Fields as ambient, infrastructural AI environments

A field is what arises when stacks and centaurs become so common and so integrated into institutional life that they form an ambient environment. In a field, AI is not experienced as a discrete tool that one occasionally uses. It is part of the background conditions under which decisions are made, stories are told, and power is exercised.

This can take many forms. In some cases, fields involve explicit coordination: fleets of centaur teams in hospitals, schools, or companies, all connected to shared data and oversight systems. In other cases, the field is more diffuse: personalization and filtering systems that shape what people see, hear, and think across many platforms, informed by model-mediated inferences. Over time, fields can influence everything from hiring and lending to news consumption and social norms.

In a mature field, opting out becomes difficult. The choice is not simply to use or not use a particular interface. The way jobs are structured, the way services are delivered, and the way institutions operate may all assume the presence of AI mediation. As a result, people who lack access to centaur capabilities may be disadvantaged, and those who are uncomfortable with the technology may feel pressure to accept it anyway.

Fields magnify both the strengths and weaknesses of underlying centaur patterns. Where centaurs are well formed, fields can support better care, more accurate decisions, and more humane institutions. Where centaurs are weak or misaligned, fields can entrench bias, normalize surveillance, and make it easier to treat people as numbers. Once a field is in place, its effects become hard to see and harder to reverse.

For this reason, the move from centaur to field is politically and ethically charged. It is not simply scaling up what works. It is building an environment that will shape what future humans can become. That shift should not be undertaken lightly or under the banner of inevitability.

2.5 Why the arrows between levels are not automatic

The conventional story of technology encourages the assumption that movement from models to stacks to centaurs to fields is a natural progression. First researchers build models. Then companies productize them into stacks. Users adopt them and become centaurs. Eventually, the tools become ubiquitous and form fields. On this view, the only questions are how fast the progression occurs and who profits.

This paper rejects that assumption. The arrows between levels are not automatic. They are contingent on choices, incentives, and values. A model can remain a research artifact, never deployed beyond a lab. A stack can be tightly constrained to a narrow domain, without encouraging centaur patterns at all. A society can decide not to let certain centaur practices scale into fields if they threaten core commitments to human dignity.

There are several reasons to insist on this contingency. First, each transition introduces new kinds of risk that are not simply extensions of what came before. A model's errors are limited to its immediate outputs. A stack's misbehavior can affect thousands of users. A centaur's failures can have moral and relational consequences that cannot be captured by technical metrics alone. A field's distortions can change entire cultures.

Second, the transitions are shaped by institutional and cultural decisions. Whether a model becomes a stack depends on a business case, a regulatory environment, and a risk calculus. Whether a stack encourages centaur patterns or treats users as passive endpoints depends on interface design, training, and incentives. Whether centaur practices are allowed to form a field depends on political debate, public consent, and competing visions of the good.

Third, treating the arrows as automatic undermines human responsibility. If it is taken for granted that "technology will spread anyway," then there is little motivation to set conditions, draw red lines, or say no. The language of inevitability becomes a cover for inaction and for the consolidation of power by actors who benefit from rapid, uncritical expansion.

By contrast, recognizing that each arrow is a gate invites a different attitude. Before moving from one level to the next, we can ask what must be true for the transition to be morally acceptable. What controls and understandings are needed before a model can be safely embedded in a stack? What kind of formation is required before a person can responsibly become a centaur in a high-stakes domain? What kinds of oversight, reversibility, and infrastructural love must be in place before centaur patterns are scaled into fields?

The answer to these questions cannot be given once and for all, but the questions themselves are non-negotiable. They are a way of insisting that human beings are more than raw material for systems, and that technological evolution is not something that happens to us but

something we participate in. In the sections that follow, the focus will narrow to the centaur level, since that is where individual responsibility and structural design meet most directly.

3. Defining the Centaur: Beyond Power Users and Tools

The centaur has often been used as a convenient metaphor for human–AI collaboration, borrowed from chess commentary and popularized as shorthand for “humans plus machines working together.” In this light, it can sound like branding rather than a serious concept. To serve as a real guide for design and ethics, however, the centaur cannot remain a vague image. It must be defined as a specific role and pattern of practice, with clear contrasts to adjacent ways of using AI.

This section develops such a definition. It begins by treating the centaur as a role that can be occupied well or badly, rather than as a mythic creature or a generic metaphor. It then distinguishes centaurs from three nearby patterns that are easy to confuse with it: ordinary tool users, domesticated “pets” of opaque systems, and operators who execute scripts without true engagement. From there, it explores the idea of joint agency, arguing that centaur patterns only become meaningful when humans retain moral responsibility even as they lean on their stacks. Finally, it turns inward, examining the inner life of a centaur: the habits of discernment, doubt, and conscience that either flourish or wither under the influence of powerful systems.

Taken together, these elements frame the centaur not as a marketing slogan but as a demanding vocation. To be a centaur is to inhabit a particular kind of relationship with technology, one that can be recognized, taught, supported, and held accountable.

3.1 Centaur as a role, not a metaphor

If the centaur is only a metaphor, it will drift toward whatever people want it to mean. One team may invoke it to celebrate speed and efficiency. Another may use it to claim that “humans are still in the loop” even when they are little more than button-pressers. Without a more precise account, the term risks becoming a way to avoid concrete questions rather than confront them.

Treating the centaur as a role changes the stakes. A role is something a person steps into, with associated expectations, competencies, and responsibilities. There can be novices and experts, good and bad examples, training curricula and codes of conduct. Roles can be institutionalized in job descriptions, professional standards, and regulatory frameworks. They can be critiqued and reformed.

As a role, the centaur has at least three defining characteristics. First, it is stable over time. It is not a one-off interaction with an AI, but a sustained pattern in which a person routinely works

through a particular stack. Second, it is reflexive. The centaur role includes the capacity to reflect on how the stack is shaping one's attention, judgments, and relationships. Third, it is answerable. The person occupying the role understands that others will rightly hold them responsible for the outcomes of the joint work, not just for their keystrokes.

This way of framing things invites a simple question whenever AI is introduced into a domain: are we asking people to become centaurs, or something else? If the answer is "centaur," then it becomes necessary to specify the role: what training is provided, what authority is conferred, what safeguards exist, and how success or failure will be judged. The centaur stops being a vague promise and becomes a concrete social position.

3.2 Distinguishing centaurs from tool users, pets, and operators

To see the centaur more clearly, it helps to hold it next to three patterns that often masquerade as collaboration but are structurally different.

A tool user is someone who employs AI occasionally or superficially, without deep integration into their way of working. They might paste text into a chat interface, copy useful fragments, and move on. The tool may save time or spark ideas, but the user's basic habits of thought, responsibility, and practice remain largely intact. There is no stable joint creature, only intermittent assistance.

A pet is a human who has become dependent on an opaque system that quietly runs their life. The system nudges their choices, filters their information, and automates their routines in ways that are hard to see and harder to contest. The person may feel cared for, but they rarely understand how their environment is being shaped. Their agency narrows, even as their comfort and convenience grow. They become domesticated to the system's incentives.

An operator is someone who sits in front of AI-driven systems but has little real authority. They execute scripts, follow prompts, and click to approve outputs that they are neither trained nor empowered to question. The organization treats them as a buffer for liability but does not give them the autonomy or time required to exercise genuine judgment. When things go wrong, they are blamed for not stopping a machine they were never trusted to truly oversee.

Centaurs differ from all three. Unlike tool users, they have a deep, ongoing relationship with a specific stack that shapes their work in a sustained way. Unlike pets, they maintain a posture of active interpretation rather than passive consumption; they treat the system as a partner to be understood, not a caretaker to be obeyed. Unlike operators, they possess both the knowledge and the institutional backing to override, slow down, or refuse the stack when conscience or expertise demand it.

These distinctions matter because fields can be built around any of these patterns. A world of centaurs would look very different from a world of pets or operators, even if the underlying models were identical. Naming the differences is a step toward choosing which future to cultivate.

3.3 Joint agency and the retention of moral responsibility

One of the central questions raised by centaur practice is simple to ask and hard to answer: who is acting when a centaur acts? When a hybrid of human and stack issues a medical recommendation, a legal strategy, or a policy draft, it is tempting to attribute agency to an abstract “system” and leave it at that. Yet this abstraction hides the moral reality of the situation.

From an ethical perspective, it is crucial that the human partner in a centaur retains moral responsibility for the actions taken. The stack may be complex, and the path from input to output may be opaque, but it is still the human who chooses to deploy the stack in a particular context, to rely on its suggestions, and to present its outputs as authoritative. Even when organizational cultures encourage blame-shifting to “the algorithm,” the centaur role insists that the person remains answerable.

This does not mean that the human is alone. Joint agency is real. The stack constrains and enables what the human sees and considers. It may surface options the human would never have imagined, or filter out concerns that would otherwise have demanded attention. It can accelerate good decisions and amplify bad ones. In this sense, the stack participates in agency, even though it is not a moral agent in the same way a person is.

Recognizing joint agency means designing systems and institutions that respect both sides. On the technical side, stacks should be instrumented to expose enough information about their behavior that human partners can make informed judgments. On the human side, training and norms should prepare centaurs to question, resist, and supplement the stack, rather than treating its outputs as inevitable or infallible.

Importantly, the retention of moral responsibility is not just a burden. It is also a safeguard against subtle forms of dehumanization. If a person can say, in good faith, “I am responsible for what happens when I use this system,” they are more likely to protect others from harm, to notice when something feels wrong, and to insist on changes. If responsibility is dissolved into a diffuse “we” or attributed to the machine, those impulses are dulled.

Joint agency, then, is not a way to evade blame but a way to acknowledge that powerful tools do change how responsibility is exercised. The centaur role calls for practices that keep the human side of that joint agency alive, active, and accountable.

3.4 The inner life of a centaur: discernment, doubt, and conscience

So far, the centaur has been described from the outside: roles, responsibilities, patterns of action. There is also an interior dimension that is equally important. Working every day with a powerful stack does not just change what a person can do; it changes how they think and feel. It shapes their sense of competence, their tolerance for ambiguity, their relationship to doubt and conviction.

A healthy centaur cultivates discernment. Discernment is the ability to sense when something is off even if one cannot immediately articulate why. It involves noticing subtle mismatches between the stack's outputs and lived reality, between what the system suggests and what previous experience or moral intuition would recommend. Discernment grows with practice. The more a centaur reflects on past interactions, learns from near misses, and seeks feedback from others, the more tuned their sense of fitness becomes.

Doubt plays a complicated role. On one hand, unthinking deference to the stack is dangerous. It leads to automation bias, where people override their own expertise to follow machine suggestions. On the other hand, constant suspicion can paralyze action and erode trust in any tool. The inner life of a centaur involves learning when to lean on the stack with confidence, when to pause, and when to seek independent verification. This calibrated doubt is not a technical skill alone; it is tied to temperament, humility, and courage.

Conscience is the deepest layer. A centaur must be able to feel when a proposed course of action violates their sense of what is right, regardless of how efficient or profitable it appears. Conscience can be sharpened or dulled by repeated choices. If a centaur regularly suppresses discomfort in order to meet metrics or please superiors, the discomfort will eventually fade. If they instead honor those signals, even when it is costly, their conscience remains sensitive.

The design of stacks and the culture of institutions directly affect this inner landscape. Interfaces that present outputs as flawless and final discourage discernment and doubt. Dashboards that reward speed over care train people to ignore their misgivings. Environments that punish refusal make it risky to act on conscience. Conversely, tools that surface uncertainty, invite questions, and make it easy to annotate or correct outputs support the inner work of becoming a good centaur.

Ultimately, the centaur is not just a technical configuration but a spiritual and psychological condition. It involves learning to live with an always-available partner that is neither a simple instrument nor a rival mind, but something in between. The quality of that life depends as much on interior habits as on external design. If exploration of the centaur is to become a central theme in the next evolutionary step, it must include attention to these quiet, interior shifts, not only to workflows and architectures.

4. Evolutionary Forks: Pets, Oligarchs, or Centaurs

Once models exist and stacks are deployed, the question is not only what the technology can do, but what kinds of human roles and power structures crystallize around it. The same technical substrate can support very different social arrangements. Three broad patterns are particularly salient: humans as pets of opaque systems, oligarchs using AI as an exoskeleton for domination, and centaurs cultivating shared power without erasing human depth.

These are not mutually exclusive caricatures. Elements of each can coexist within the same society, organization, or even person. Nonetheless, they represent distinct attractors in the space of possibilities. Which attractor dominates will shape not only immediate outcomes but the character of any future field in which AI becomes infrastructural.

This section describes these three patterns and argues that serious exploration of the centaur pattern is the only humane next step. It then examines what is likely to happen if centaur development is neglected in favor of pet-like dependency or oligarchic augmentation.

4.1 The pet pattern: humans as front-ends for hidden decisions

In the pet pattern, humans are still physically present and apparently active, but critical decisions are made elsewhere. The systems that structure daily life are opaque, paternalistic, and tuned to maximize engagement, convenience, or compliance. People are given sleek interfaces, reassuring notifications, and personalized recommendations, but the underlying logic is hidden behind proprietary walls or explained in language too vague to act upon.

This pattern can emerge without explicit malice. Designers genuinely want to reduce friction. Personalized feeds keep users entertained. Smart assistants simplify logistics. Automated scoring systems promise efficiency in hiring, lending, and security. Over time, however, these conveniences accumulate into a dense mesh of algorithmic mediation. The human becomes the visible tip of a decision iceberg, tapping and swiping on options preselected by systems they do not control.

The pet pattern is characterized by three features. First, dependence: as more tasks are automated, people lose the skills and confidence to act without the systems. Second, opacity: the criteria by which the systems decide are difficult to inspect, challenge, or change. Third, misaligned incentives: many of the systems are designed to serve the interests of their owners rather than the long-term flourishing of users.

From the outside, this arrangement can look benign. Pets are fed, entertained, and sometimes deeply loved. From the inside, it risks producing a subtle form of captivity. Agency narrows to choosing among prefiltered options. Resistance requires swimming against the current of every

optimized interface. When this pattern scales, it shapes a field in which human beings are managed rather than genuinely collaborated with.

4.2 The oligarch pattern: AI as an exoskeleton for concentrated power

In the oligarch pattern, AI serves as an exoskeleton for a relatively small number of actors who already possess significant power. Governments, corporations, or wealthy individuals use advanced stacks to amplify their reach, surveillance capacities, and ability to shape narratives. The focus is not on broad empowerment but on strategic advantage.

Here, the defining traits are asymmetry and centralization. A limited group has access to the best models, the richest data, and the most integrated tools. They can analyze populations, simulate scenarios, target messaging, and adjust policies with unprecedented precision. Others interact with AI primarily through interfaces designed and controlled by these central actors. Even when tools are made widely available, their most potent capabilities may remain reserved for those who can pay or who sit inside privileged institutions.

This pattern can coexist with a rhetoric of democratization. Public-facing tools and partnerships are highlighted, while the internal use of AI for control, extraction, or manipulation remains obscure. The net result is not a flat landscape of empowered centaurs but a steep hierarchy. At the top, AI-augmented elites operate with extended sensory and computational apparatus. Below, users navigate an environment increasingly shaped by systems they cannot meaningfully influence.

The oligarch pattern is particularly dangerous when combined with the pet pattern. When those who design and own the systems also control the surrounding field, pets may be deeply loyal to arrangements that systematically disadvantage them. Small groups can steer large populations while maintaining a veneer of choice and personalization. In such a world, the possibility of authentic centaur agency shrinks, not because the technology forbids it, but because the structures of power make it rare and risky.

4.3 The centaur pattern: shared power with preserved human depth

The centaur pattern presents a different trajectory. Instead of using AI primarily to manage or dominate others, it uses AI to deepen individual and collective agency. The goal is not to eliminate human limitations, but to extend human capacities in ways that preserve the conditions for moral and spiritual growth.

In a centaur pattern, individuals or teams work closely with stacks while retaining real authority and responsibility. They understand enough about the systems they rely on to question them. They are trained and encouraged to override, suspend, or reconfigure the tools when

circumstances demand. Their expertise is not sidelined by automation but refined through interaction with powerful models.

Shared power is central. The stack contributes speed, recall, and pattern recognition. The human contributes judgment, context, and care. Neither side is fully replaceable by the other. Workflows are designed so that decisions that significantly affect lives always pass through human conscience. Metrics emphasize not only throughput and profit but quality of understanding, fairness, and long-term impact.

Preserving human depth means more than keeping humans “in the loop.” It means structuring education, work, and culture so that people continue to develop virtues such as patience, courage, honesty, and compassion even in highly augmented environments. It requires tools and institutions that invite reflection rather than constant distraction, that reward integrity rather than mere compliance, and that allow for genuine disagreement between human and machine.

Centaur patterns can exist at many scales. A single clinician working with a diagnostic system can be a centaur. So can a newsroom where journalists collaborate with AI-based research and verification tools. So can a small company that uses language models to explore possibilities while still anchoring decisions in shared values. What unites these cases is a commitment to using AI to amplify the best of human capacities while guarding against the erosion of responsibility and care.

4.4 Why centaur exploration is the only humane next step

If one accepts that powerful AI systems are here to stay, the question is not whether humans will adapt, but how. Among the three patterns described, only the centaur pattern treats the combination of human and AI as an opportunity for ethical evolution rather than mere control or comfort.

Exploring the centaur pattern is humane for several reasons. First, it takes seriously the moral status of human beings. Rather than treating people as objects to be optimized around, it treats them as agents whose freedom and growth matter. Second, it acknowledges the real benefits of AI capabilities without surrendering to the illusion that those capabilities can or should replace human judgment in matters of value. Third, it provides a framework for designing institutions that do not quietly train people into passivity.

Centaur exploration also has a prudential dimension. Societies composed mainly of pets and oligarchs are unstable. Resentment builds when many feel managed and few bear responsibility. Errors and abuses compound when decision-making is opaque and unchallengeable. By contrast, widespread centaur practice distributes both power and accountability more broadly.

It equips more people to notice and contest harmful uses of technology, and to propose better ones.

Finally, focusing on centaurs aligns technological development with a vision of the good that extends beyond efficiency. It allows discussions of beauty, love, and meaning to enter the design of systems, not as ornamental afterthoughts but as criteria for success. If the next evolutionary step is to be more than a race toward ever finer control, it must include a serious attempt to understand what kind of joint creatures we are becoming and what kinds we wish to be.

4.5 Skipping the centaur: structural consequences for future fields

What happens if centaur exploration is neglected? Suppose the world slides directly from stacks to fields, with little attention to cultivating robust, responsible centaur roles. The likely outcome is a field shaped disproportionately by pet and oligarch patterns.

In such a field, the majority of people interact with AI as a convenience layer and background presence, not as a partner in joint agency. Their skills gradually adapt to interfaces, not to the underlying realities those interfaces mediate. They may become highly adept at prompting systems or navigating recommendation flows, but less practiced at understanding how decisions are made or how to intervene when something goes wrong.

At the same time, those who control the underlying stacks gain a double advantage. They not only possess superior technical tools but also operate in a landscape where most others lack the habits and structures needed to resist or negotiate. Feedback from the field is filtered through layers of automated systems, making it easier to ignore dissent or to treat it as noise. The gap between those who shape the environment and those who inhabit it widens.

Structurally, this creates brittle dependencies. If critical infrastructure is built around AI-mediated processes without strong centaur participation, failures can cascade in unpredictable ways. Bugs, adversarial attacks, or misguided policies can ripple through systems that few fully understand. When something breaks, there may be too few people with the combination of technical and moral insight needed to diagnose the problem and propose humane fixes.

Skipping the centaur also impoverishes culture. Without lived examples of healthy human–AI cooperation, stories and norms will be drawn from extremes: fearful tales of machine domination or enthusiastic celebrations of total automation. The subtle work of learning how to live well with powerful tools will be underrepresented. Future generations may inherit a field that is technologically sophisticated but spiritually thin.

In short, the structural consequences of neglecting centaur development are profound. They include concentrated power, widespread dependency, fragile systems, and a narrowing of the moral imagination. By contrast, investing now in understanding and nurturing centaur patterns

offers a chance to build fields that, if they eventually come, bend around human dignity rather than eroding it.

5. Empirical Centaurs: Where Joint Agency Already Exists

So far the centaur has been treated as a conceptual role and a normative ideal. Yet the patterns that define it are already emerging in practice. Across domains as varied as writing, research, medicine, education, and the arts, humans are working daily with AI stacks in ways that go beyond intermittent tool use. These early centaurs differ in quality and self-awareness, but they provide concrete examples of what joint agency looks like and what can go wrong.

This section surveys several such patterns. It begins with knowledge work, where language models have become ubiquitous and where centaur behaviors are increasingly visible. It then turns to clinical, educational, and creative settings, where stakes and sensibilities differ but the underlying dynamics are similar. The distinction between strong and weak centaur practices is introduced as a way of separating careful, responsible hybrids from shallow dependence. Finally, lessons are drawn from early adopters and implicit centaur cultures that have formed without necessarily using the name.

The goal is not to provide exhaustive empirical data but to sketch a landscape. These examples show that the centaur is not a distant possibility but a live phenomenon. They also show that centaurs can be formed by accident or by design, and that intentional cultivation makes a difference.

5.1 Emerging patterns in writing, research, and knowledge work

In writing and research, centaur patterns are becoming common. Knowledge workers use language models to outline arguments, draft emails and reports, summarize documents, and explore unfamiliar topics. Initially this often takes the form of sporadic tool use: a query here, a paragraph completion there. Over time, however, some people settle into stable routines where nearly every significant piece of written work is co-produced with an AI stack.

Several characteristic behaviors emerge. One is iterative co-drafting. The human sketches an idea, the model elaborates, the human trims and redirects, the model suggests alternatives, and so on. Authorship becomes genuinely joint. Another is exploratory questioning. Researchers use models to brainstorm hypotheses, identify blind spots, and rehearse objections before committing to a line of argument. The stack serves as a sparring partner, not just a text generator.

There are also patterns around memory and retrieval. Knowledge workers build personal corpora of notes, papers, and archives that are indexed by AI tools. Over time, they come to rely

on these stacks as extensions of their own memory, querying them not only for facts but for earlier versions of their own thinking. The centaur’s “mind” thus spans both human recollection and machine-augmented retrieval.

These practices differ sharply from one-off uses of AI to polish grammar or fix formatting. A writer who occasionally asks for a synonym remains largely independent of the tool. A writer who develops a daily practice of co-drafting with a model, reviewing and revising outputs critically, and building a shared corpus for retrieval is inhabiting a centaur role. Their cognitive life is entangled with the stack in an ongoing way.

This entanglement is neither good nor bad in itself. It can lead to deeper thinking and more careful expression if the human partner remains engaged, skeptical, and responsible. It can also lead to shallower work if the model’s outputs are accepted uncritically. What distinguishes strong centaurs from weak ones in this domain is not the amount of AI usage but the quality of attention and judgment applied to it.

5.2 Clinical, educational, and creative centaurs

Beyond knowledge work, centaur patterns are emerging in domains where human lives and development are directly at stake.

In clinical settings, decision support systems have long existed in simpler forms. With more capable models, clinicians can now consult tools that synthesize patient histories, medical literature, and guidelines into suggestions for diagnosis and treatment. When used wisely, these systems become parts of a centaur practice. The clinician retains final authority, uses the tool to surface possibilities, checks its recommendations against experience, and involves patients in decisions. The stack serves as a resource and a second opinion, not as a hidden arbiter.

The risks are clear. Under time pressure, clinicians may begin to rely too heavily on recommendations, especially when those recommendations are framed in confident language. Over time, skill can erode if practitioners are not deliberate about maintaining their own diagnostic reasoning. Strong clinical centaurs actively resist this erosion. They treat the system as an assistant that must pass through the filter of professional judgment and the obligations of care.

In education, teachers experiment with AI-assisted lesson planning, personalized exercises, and feedback. A teacher who uses a stack to generate drafts of materials, then adapts them to the needs of particular students and the realities of the classroom, is forming a centaur pattern. The stack extends the teacher’s capacity to differentiate instruction and to find examples or explanations they might not have time to prepare alone. The teacher’s role remains central: interpreting outputs, contextualizing content, moderating pace, and caring for the emotional and social life of the class.

If the balance shifts too far toward automation, however, education can drift toward the pet pattern. Students receive personalized content generated and graded by systems, while teachers become supervisors of software rather than formative presences. Centaur exploration in education thus includes not only new tools but clarification of what uniquely human aspects of teaching must be protected and emphasized.

Creative fields also show early centaur behaviors. Artists, musicians, and designers use generative systems to explore variations, test compositions, and overcome blocks. Some treat these systems as mere tools for texture or polish. Others develop deep collaborative practices, where the stack becomes a kind of studio partner, offering unexpected directions that the human then curates and develops. The resulting centaur can have a distinct style that neither component would produce alone.

In all these domains, centaurs operate under different constraints and values. Clinical centaurs are bound by do-no-harm principles, privacy obligations, and regulatory frameworks. Educational centaurs are guided by developmental goals and care for students' autonomy. Creative centaurs navigate questions of authorship, originality, and meaning. The common thread is that centaur practice involves sustained, reflective cooperation with a stack in service of human ends that go beyond efficiency.

5.3 Strong versus weak centaur practices

Not all centaurs are equally robust. Some practices deserve the name only in a thin sense: the human and the stack are linked, but the human's agency is fragile and easily overridden. Others exhibit a depth of joint agency that makes them worth emulating. The distinction between strong and weak centaur practices helps make this visible.

Weak centaur practices share several traits. The human partner treats the stack as authoritative by default, overriding it only when it obviously fails. Prompts are reused mechanically without much thought about what they imply. Outputs are lightly edited and forwarded on, often under time pressure. There is little reflection on how reliance on the stack is changing skills, relationships, or ethical responsibilities. The arrangement functions, but it is brittle. If the stack's behavior shifts or its incentives change, the human may not notice until harm has occurred.

Strong centaur practices look different. The human partner understands the limits and failure modes of the stack and routinely designs workflows that compensate for them. There is a habit of double-checking in high-stakes contexts, and of seeking independent corroboration when outputs surprise or unsettle. The centaur maintains a living sense of which tasks are safe to delegate, which require intensive human oversight, and which should not be mediated by AI at all.

Training and institutional support distinguish strong from weak centaurs. Where organizations invest in teaching people how the systems work, what their biases are, and how to challenge them, stronger practices are more likely to emerge. Where metrics reward speed and volume over judgment, weak practices proliferate. Individual character also matters. People who bring humility, curiosity, and a sense of responsibility to the partnership tend to grow into stronger centaurs over time.

Another marker of strong practice is reversibility. Strong centaurs maintain the ability to do their work without the stack, at least in degraded form. They continue to practice core skills. This is not always possible at full scale, but the aspiration matters. It keeps the human from becoming interchangeable with an interface and preserves a baseline of independent competence.

5.4 Lessons from early adopters and implicit centaur cultures

Early adopters of AI-assisted work offer lessons about how centaur cultures form and what they need to thrive. Many of these centaurs did not set out to occupy a new role. They simply began experimenting with new tools and, over time, found themselves in stable partnerships with them. Their experiences point to patterns that can be nurtured intentionally.

One lesson is that centaur roles emerge more easily when individuals are given room to experiment and reflect. Those who have time to try different workflows, to compare outcomes with and without AI assistance, and to discuss experiences with peers develop more nuanced practices. Environments that treat AI as a rigidly specified productivity tool, with little tolerance for exploration, tend to produce weaker centaurs.

A second lesson is that peer communities matter. Informal exchanges about prompts, pitfalls, and surprising behavior help people build a shared understanding of how to work with stacks responsibly. Over time, these communities can develop norms and informal codes of conduct that go beyond what official guidelines provide. Where such communities are encouraged, centaur cultures can form that are resilient and self-correcting.

A third lesson concerns leadership. In organizations where leaders model careful, reflective use of AI and speak openly about both its promise and its dangers, centaur practices are more likely to align with ethical commitments. When leaders focus only on efficiency gains and avoid discussing responsibility, people receive a different message: the goal is to maximize throughput, and concerns about erosion of skill or conscience are secondary.

Implicit centaur cultures already exist in some fields that have long histories of working with complex instruments. Pilots, for example, operate as centaurs with autopilots and flight management systems, retaining responsibility while relying on automation. Good aviation cultures emphasize checklists, cross-checking, and the authority of humans to disconnect systems when necessary. While AI tools are different in kind, these cultures demonstrate that it

is possible to institutionalize respect for human judgment within technologically dense environments.

The main lesson from early centaurs is simple. Joint human–AI agency is not hypothetical. It is already shaping work, thought, and care. Where it is nurtured with attention to responsibility, skill, and love, it can become a path of humane evolution. Where it is left to chance or captured by narrow incentives, it can slide toward pet-like dependence or oligarchic control. Recognizing and learning from these early patterns is therefore not a luxury but a necessity for anyone concerned with the future of fields built on AI.

6. Norms and Oaths: What a Good Centaur Refuses

If centaurs are to be more than accidental byproducts of tool adoption, they will need norms. Not just informal habits or scattered best practices, but clear statements of what this role allows and what it forbids. In high-stakes domains, such norms have often taken the form of oaths and codes: explicit promises about how power will be used and what lines must never be crossed.

For centaurs, the central danger is not that they will lack capability, but that they will gradually surrender conscience and skill to the stack. The temptation is to let the machine not only help them see more, but also decide for them what is worth seeing and what is acceptable to do. A good centaur refuses that temptation. They use AI to clarify, not to anesthetize; to see more clearly, not to feel less accountable.

This section explores five elements of a centaur ethic. It begins with the distinction between offloading conscience and using AI to deepen perception. It then insists on a norm against hiding behind “the model said” as an excuse. From there it turns to design: how tools and workflows can either accelerate or resist deskilling and moral atrophy. A draft centaur oath is proposed as a concrete articulation of commitments and prohibitions. Finally, the section argues that formation, not just technical training, is necessary if these norms are to take root.

6.1 Offloading conscience versus using AI to see more clearly

The most subtle failure mode for centaurs is not dramatic harm, but quiet numbing. A person begins by using a stack to help them understand complex situations, to surface overlooked aspects, and to check their own biases. Over time, the same system can become a way to avoid the discomfort of moral conflict. When faced with a hard choice, the centaur asks the model what to do and follows its suggestion. The pain of ambiguity is lowered; the burden of responsibility feels lighter.

At that point, the stack has become a surrogate conscience. It offers tidy resolutions to dilemmas that should hurt. The centaur no longer wrestles with tradeoffs because the interface

presents a clean recommendation. The danger is not that the model is evil, but that the human stops doing the inner work that conscience requires.

A good centaur draws a sharp line here. They treat the stack as an instrument for seeing more of the situation, not for deciding what is right. When confronted with a difficult case, they might use AI to map stakeholders, list possible consequences, identify relevant precedents, or articulate arguments for and against each option. These are ways of using AI to see more clearly. The decision itself, however, remains a matter of human judgment.

This distinction can be practiced. A centaur can develop the habit of asking models questions like “What am I not seeing?” or “What arguments could be made against this plan?” rather than “What should I do?” They can resist prompts that push models into a pseudo-moral voice. They can treat the moment of decision as a point where the stack goes quiet and the human, perhaps in conversation with other humans, must answer.

6.2 Refusing to hide behind “the model said”

When things go wrong, there will be pressure to distribute responsibility so thinly that no one feels accountable. In centaur settings, one of the most tempting evasions is the phrase “the model said.” It suggests that the human simply followed the output of a system and therefore cannot be blamed. This pattern already appears in simpler algorithmic contexts and is likely to intensify as models grow more capable.

A core centaur norm must be to refuse this refuge. The presence of a powerful stack does not eliminate agency; it complicates it. The human partner chose to rely on the system, chose how to prompt it, chose which outputs to trust, and chose how to present them to others. Even under strong institutional pressure, a centaur ethic insists that these choices are morally significant.

Practically, this means several things. First, centaurs should use language that keeps their own agency visible. Instead of saying “the model says X,” they can say “the model suggests X; I have checked it against what I know and I think it is appropriate because...” This keeps the human voice in the loop and reminds others that the centaur is more than a transmission channel.

Second, centaurs should resist organizational cultures that treat model outputs as mandatory. If policies or informal expectations penalize those who deviate from model recommendations, the ability to retain responsibility is compromised. A centaur code should therefore include the right, and sometimes the obligation, to override the stack without fear of automatic reprisal.

Third, centaurs must be ready to explain their reasoning. Hiding behind “the model said” is often a way to avoid the hard work of articulating why a decision makes sense. By contrast, a

centaur willing to give reasons signals respect for those affected. They are saying, in effect, “I did not simply forward an output; I thought about you.”

6.3 Designing against deskilling and moral atrophy

Ethics is not only about personal resolve; it is also about structure. Even the best intentions can erode if tools and workflows are designed in ways that steadily deskill users and dull their moral senses. For centaurs, an important part of the oath must therefore concern how systems are built and adopted.

Deskilling occurs when tasks that used to require human expertise are automated in such a way that the underlying knowledge fades. If a clinician always follows model suggestions and seldom practices independent diagnostic reasoning, their skills will degrade. If a teacher relies entirely on AI for lesson design and feedback, their intuition about pacing, explanation, and student struggle may weaken.

Design can push against this. Systems can be built to expose intermediate reasoning steps rather than only final recommendations, encouraging users to engage with the process. Workflows can include regular exercises or cases where the centaur works without AI, maintaining a baseline of independent competence. Metrics can reward thoughtful oversight and improvement, not just raw throughput.

Moral atrophy is related but deeper. It happens when repeated choices train a person to ignore their discomfort or to treat others as abstractions. Interfaces that hide the human faces behind data, or that present morally significant decisions as trivial clicks, contribute to this erosion. A centaur ethic calls for designs that keep consequences vivid. For example, in high-stakes contexts, systems might periodically remind centaurs of previous cases where similar decisions had serious impacts, or provide space for reflection before irreversible actions.

Centaurs themselves can participate in design choices. They can ask, before adopting a new tool, how it will affect their skills over time. They can advocate for features that support learning and reflection. They can push back against configurations that turn them into passive supervisors of automation. In this way, the norm against deskilling and moral atrophy becomes part of both personal practice and institutional infrastructure.

6.4 A draft centaur oath: commitments and prohibitions

To make these norms concrete, it is helpful to sketch a draft centaur oath. This is not a final text, but a starting point for reflection and adaptation in different domains.

A centaur might promise something like the following:

- I acknowledge that working with AI systems amplifies my power to affect others for good or ill. I accept responsibility for how I use this power.
- I will use AI to help me see more clearly, not to escape the discomfort of hard choices. I will not delegate my conscience to a model.
- I will not present model outputs as authoritative without understanding their limits. When stakes are high, I will seek corroboration and exercise independent judgment.
- I will not hide behind phrases like “the model said” to evade accountability. I will own my decisions, including those shaped by AI assistance.
- I will resist designs and incentives that deskill me or dull my concern for others. Where possible, I will advocate for tools and workflows that sustain human expertise and empathy.
- I will protect the privacy, dignity, and non-disposability of the people affected by my work. I will not use AI to manipulate or dehumanize them, even when such use is permitted or profitable.
- When I recognize that my centaur practice is harming my character or my relationships, I will seek counsel, adjust my use of AI, or step back from certain tasks.

This oath also implies prohibitions. A centaur commits not to use AI to fabricate convincing lies about real people, to bypass consent, to create systems whose main purpose is addiction or coercion, or to design fields that leave others with no meaningful alternatives. Some of these prohibitions will need to be spelled out in law and policy. Others will remain matters of professional and personal integrity.

The point of an oath is not perfection. It is to set a direction and a public standard that can be used to educate, to correct, and, when necessary, to judge. Centaur oaths will likely vary across professions, but they can share a common core: commitment to responsibility, clarity, and care in the use of AI.

6.5 The role of formation, not just training

Technical training can teach people how to prompt models, interpret outputs, and navigate interfaces. It cannot by itself produce the character needed to live the centaur role well. For that, something deeper is required: formation.

Formation refers to the long, often slow process by which people develop habits, virtues, and sensibilities. It happens through mentorship, community, repeated practice, and the stories a

culture tells about what is admirable. In medicine, formation includes learning not only anatomy and pharmacology but bedside manner, humility in the face of uncertainty, and the courage to advocate for patients. In centaur practice, formation must include learning how to live with a powerful, always-available partner without surrendering one's humanity.

This involves several elements. First, centaurs need communities where they can speak honestly about the pressures they face, the temptations to cut corners, and the subtle ways their inner life is changing. These communities can be professional associations, teams, or informal circles, but they must be places where both technical and moral questions can be discussed without pretense.

Second, centaurs benefit from guided exposure to difficult cases. Just as medical trainees learn from case studies that force them to confront dilemmas, centaur formation can include narratives where AI use went wrong and where it went well. Reflecting on these stories helps build moral muscle and anticipatory wisdom.

Third, centaur formation requires exemplars. Seeing someone live the role well—balancing reliance and skepticism, using AI to enhance care rather than escape it, accepting responsibility when things go wrong—teaches more than any abstract principle. Organizations that value centaurs must therefore highlight and support such people, not only as high performers but as models of integrity.

Finally, formation reminds us that centaur ethics cannot be reduced to a checklist. It is about becoming a kind of person who can be trusted with hybrid agency. Training answers the question “Can you do this?” Formation answers the question “Should you be the kind of person who does this?” A humane future for AI fields depends on both.

7. Infrastructural Love: Constraints on Centaur and Field

Up to this point, the focus has been on roles, practices, and architectures. This section asks a more uncomfortable question: what, if anything, should constrain these systems at the deepest level? If centaurs are the next evolutionary step and fields are the environments they may eventually inhabit, what principle should govern their design beyond efficiency, safety, or legality?

One answer is love. Not love in a sentimental sense, but love expressed as non-disposability and care built into structure. Infrastructural love is what it looks like when the systems that scaffold daily life make it materially hard to treat people as expendable or invisible. It is a commitment to designing AI-mediated environments in which it is easier to honor human worth than to trample it.

This section explores what such love would mean for centaurs and fields. It begins by defining love at scale in structural terms. It then introduces the “Beautiful” constraint, a concrete way of testing whether deployments are aligned with this vision. From there it considers how interfaces can either express or betray infrastructural love by revealing structure, uncertainty, and limits. Reversibility, kill-switches, and the right to say no are treated as practical expressions of care. Finally, it considers what would be required to embed these constraints not only in individual systems but in law, design practice, and culture.

7.1 Love at scale as non-disposability and care in structure

Love is easy to talk about at the level of individual relationships. It is harder to specify when speaking of institutions or infrastructures. At scale, love cannot be a feeling; it must be encoded as patterns that shape what is possible.

Two such patterns are central. The first is non-disposability. A system embodies love when it refuses to treat any person as a mere means to some external end. This refusal shows up in choices about who bears risk, who is sacrificed for efficiency, and whose suffering is considered acceptable collateral. In an infrastructural sense, love means designing systems so that it is difficult to profit from the quiet harm of invisible others.

The second pattern is care in structure. Care is more than emergency response or charity layered on top of neutral mechanisms. Structural care means that, in the ordinary functioning of a system, the burdens of uncertainty, error, and change are not dumped on those least able to carry them. It means that feedback channels are real, that grievances can be heard and addressed, and that people have ways to influence the systems that shape their lives.

For centaurs, love at scale implies that their augmented power is used in ways that protect, rather than exploit, those who do not share their capabilities. For fields, it implies that the ambient environment itself nudges toward mercy and mutual recognition instead of indifference and extraction. This is not achieved by slogans but by concrete constraints on design and deployment.

7.2 The “Beautiful” constraint as a deployment test

One way to hold love in view is through a constraint drawn not from technical literature but from art. Imagine that every move toward more pervasive AI had to pass a simple test: does this treat humanity in the way expressed by the song “Beautiful” by Paul van Dyk and Julia Westlin, or does it move in the opposite direction?

The song, taken as a specification, encodes a posture. It speaks of presence without domination, of holding without possession, of seeing another’s fragility and responding with gentleness rather than manipulation. Translated into design language, this “Beautiful” constraint becomes

a deployment test: any proposed centaur practice or field expansion must be evaluated against whether it tends to see people as precious, vulnerable, and worth protecting.

Concretely, the constraint could be operationalized through questions such as:

- Does this deployment make it easier or harder to treat people as non-disposable?
- Does it increase the chance that someone in distress will encounter a response shaped by compassion rather than exploitation?
- Does it support centaurs who care about others, or primarily those who seek leverage?
- If this pattern scaled, would the ambient tone of the field feel more like a protective embrace, or more like a soft cage?

These questions are deliberately qualitative. They are not replacements for quantitative risk assessments, but they supplement them by pointing to the moral atmosphere a system creates. A field built under a “Beautiful” constraint will not be perfect or free of harm. It will, however, have passed through a filter that treats tenderness and reverence for persons as non-negotiable.

7.3 Interfaces that reveal structure, uncertainty, and limits

Infrastructural love has a visible face: the interface. It is at the interface that centaurs and non-centaurs alike encounter AI systems. Interfaces can either respect users as agents capable of understanding and choosing, or they can treat them as attention sources to be steered.

Interfaces aligned with love reveal structure. They make it clear when a suggestion comes from a model, when it comes from a policy, and when it comes from a human. They indicate which tools were invoked, what kinds of data were consulted, and where important boundaries lie. This does not require full transparency of internal mechanics, but it does demand enough visibility that users can situate themselves.

Such interfaces also reveal uncertainty. Instead of presenting outputs with unfounded confidence, they indicate degrees of reliability, alternative possibilities, and known limitations. For centaurs, these signals support discernment. For ordinary users, they discourage unwarranted deference. In both cases, acknowledging uncertainty is a form of respect: it treats others as adults who can handle the fact that the world is not fully under control.

Limits are equally important. Interfaces that embody infrastructural love do not pretend that everything can be automated, predicted, or optimized. They show when the system is leaving a decision to the human because the stakes or values involved cannot be resolved mechanically. They may even include explicit “human-required” zones where the stack refuses to act without human initiative.

By contrast, interfaces that hide structure, blur the distinction between model and human voice, and suppress uncertainty encourage pet and operator patterns. They make it easier for people to slip into passive consumption or rote supervision. From the perspective of love, such designs are suspect even if they are efficient, because they quietly undermine human agency and responsibility.

7.4 Reversibility, kill-switches, and the right to say no

Care in structure demands that centaurs and communities retain the ability to step back from AI systems when necessary. Irreversible deployments, opaque dependencies, and architectures that cannot be halted without catastrophic consequences are inherently unloving. They trap people in arrangements they no longer consent to or cannot escape.

Reversibility is the opposite of fatalism. Technically, it means designing systems so that critical functions can be maintained, at least at reduced capacity, without continuous AI mediation. Socially, it means preserving practices and institutions that do not rely exclusively on algorithmic judgment. For centaurs, reversibility includes maintaining core skills and keeping alive non-augmented ways of knowing and acting.

Kill-switches are a more dramatic expression of the same principle. A system that cannot be safely turned off or heavily throttled cannot be said to serve humans in any robust sense. The presence of functioning kill-switches, overseen by accountable authorities with clear criteria for their use, is a test of whether power is truly being held in trust or has silently reversed direction.

The right to say no operates at the level of individuals and communities. People should have meaningful ways to decline participation in certain centaur practices or field dynamics without being effectively exiled from economic and social life. This does not mean that every technology can be optional in every situation, but it does mean that a field built on coerced adoption or lack of alternatives fails the test of love.

Together, reversibility, kill-switches, and the right to say no embody the recognition that humans are not tools for realizing a technological trajectory. They are the ones for whom any trajectory must ultimately answer. Systems that cannot be interrupted, questioned, or abandoned invert this order.

7.5 Embedding these constraints in law, design, and culture

For infrastructural love to be more than a personal aspiration, it must be woven into shared frameworks. Law, design practice, and culture each have a role to play.

Law can articulate minimum standards. It can require clear disclosures about AI mediation in critical decisions, mandate avenues for redress, and enforce obligations of explainability proportionate to impact. It can recognize the right of individuals to contest decisions made

about them by AI-augmented processes and to demand human review. It can also limit certain uses outright, such as systems whose primary function is manipulation or surveillance of vulnerable populations.

Design practice can internalize love as a professional norm. Designers and engineers can be trained to ask not only about usability and performance but about the non-disposability of users. Design reviews can include questions drawn from the “Beautiful” constraint. Metrics can be expanded to include signals of user flourishing, harm distribution, and preservation of agency, not just engagement and revenue.

Culture provides the stories and expectations that make such constraints either natural or strange. Narratives that glorify total automation and effortless control will make infrastructural love seem naive. Narratives that honor those who use power gently, who accept limits, and who prioritize the vulnerable will make love-shaped constraints feel like common sense. Art, music, religious traditions, and civic education all contribute to this background.

For centaurs, living within such a culture means they are not alone in their commitments. Their refusal to offload conscience, their insistence on reversibility, and their demand for interfaces that reveal structure are supported rather than undermined by the surrounding environment. For fields, it means that the invisible assumptions shaping their growth are at least partially aligned with the belief that human beings are to be cherished, not harvested.

Infrastructural love is demanding. It asks for more than safe systems or efficient workflows. It asks for fields and centaurs that, taken together, treat each person as someone who could have been the subject of a song like “Beautiful” and build around that recognition as if it were true.

8. Preparing Institutions for Centaur Cultures

If centaurs are to become more than isolated pioneers, institutions will have to change around them. At present, most organizations are implicitly optimized either for pre-AI workflows or for a shallow form of automation that treats humans as replaceable interface layers. Neither is suited to cultivating robust centaur practice. Structures, policies, and everyday habits must be reworked so that joint human–AI agency is not an awkward add-on but a supported way of working.

This section considers what such preparation might involve. It begins with education, understood broadly as the formation of future centaurs rather than mere training in tool use. It then turns to professional roles and governance structures that make responsibilities clear. Incentive systems are examined as a crucial lever for encouraging judgment instead of blind efficiency. Documentation and audit practices are considered as mechanisms for shared

responsibility in centaur teams. Finally, several institutional sketches show how these ideas might look in practice.

The common thread is that institutions cannot simply drop AI into old arrangements and hope for the best. If centaurs are to flourish, organizations must be intentionally designed to welcome and guide them.

8.1 Rethinking education: training centaurs, not just users

Education systems are already under pressure to “prepare students for AI.” Too often this is interpreted as teaching basic prompting skills or integrating generative tools into assignments without deeper reflection. If centaurs are the next evolutionary step, education should aim higher. Its task is to train people who can inhabit hybrid roles responsibly, not just operate interfaces effectively.

This requires shifts at multiple levels. Curricula need to include not only technical literacy about how models work and where they fail, but also conceptual frameworks for understanding joint agency, bias, and responsibility. Students should learn to ask what a model is good at and what it cannot see; to design workflows that check outputs rather than accepting them; and to reflect on how reliance on AI is changing their own thinking.

Pedagogy must also change. Instead of banning AI entirely or allowing unexamined use, educators can design assignments where the centaur pattern is explicit. Students might, for example, be required to document how they used a model in a project, what they accepted, what they rejected, and why. They might compare model-generated arguments with their own and analyze differences in assumptions and tone. This turns AI from a shortcut into a mirror and a partner.

At higher levels of education and professional training, centaur studios can be created: environments where teams work with AI on complex, open-ended problems under supervision. In such studios, instructors can model strong centaur practices, intervene when students slide into passivity, and make visible the ethical dimensions of hybrid work. Over time, this could become as standard as laboratory courses in the sciences.

Finally, education must preserve and deepen distinctly human capacities that risk being undervalued in an AI-rich world: patience, the ability to sit with unresolved questions, empathy, and the courage to say no. These are not optional extras. They are the traits that make someone trustworthy as a centaur. A system that trains excellent tool users but fragile consciences will not be ready for centaur cultures.

8.2 Professional roles and governance for centaur practice

Within institutions, centaurs need clear roles and governance structures. If joint human–AI agency is treated as an informal habit, responsibility will blur and conflicts will be hard to resolve. New patterns of work call for explicit recognition.

One piece of this is role definition. Job descriptions can distinguish between positions where AI is optional assistance, where it is central to the role, and where it is inappropriate. In AI-intensive roles, “centaur practitioner” or similar titles could signal that working responsibly with stacks is a core expectation. Such roles would include explicit duties: maintaining skill independent of the tools, documenting how AI is used, and participating in audits and reviews.

Governance structures must complement these roles. Institutions can establish committees or offices responsible for AI oversight, including centaur practice. These bodies would not only manage technical risks but also consider how AI use is affecting professional standards and relationships with clients, patients, students, or the public. They could issue guidelines, review incidents, and provide advice on difficult cases where centaur practice intersects with ethics or law.

There is also a need for boundary roles. AI stewards or ombuds can act as points of contact for concerns about misuse or misalignment of centaur practices. Their mandate would be to listen, to investigate, and to recommend changes, with some degree of independence from line management. Without such roles, people who notice troubling patterns may have no safe way to raise them.

Clear governance does not remove ambiguity. It does, however, give centaurs a framework within which to exercise their judgment, and it gives institutions reference points when conflicts or failures arise. Instead of improvising after the fact, organizations can build channels of accountability into centaur cultures from the beginning.

8.3 Incentives that reward judgment over blind efficiency

Incentive structures are often the most powerful forces in institutional life. If metrics reward speed, volume, or engagement above all else, then even well-intentioned centaurs will feel pressure to use AI in ways that compromise judgment. Preparing institutions for centaur cultures means rethinking what counts as success.

One approach is to balance quantitative metrics with qualitative evaluations that explicitly value careful use of AI. For example, performance reviews might include questions about how an individual has used AI to improve understanding rather than just increase throughput. Peer feedback can be solicited on whether a person is known for asking good questions of their stack, for catching subtle errors, or for resisting uses of AI that would harm others.

Error-handling policies are another lever. If an organization punishes individuals harshly for mistakes that stem from model failures, it creates an incentive to hide problems or to over-defer to systems perceived as safer politically. By contrast, if it treats honest reporting of AI-related errors as a contribution to learning, it encourages centaurs to remain vigilant and transparent.

Time allocation matters as well. Strong centaur practice requires time for reflection, double-checking, and continued skill development. Institutions that expect AI to simply compress work into fewer hours without preserving space for such activities are effectively penalizing good centaurs. Protecting dedicated time for review, learning, and discussion signals that judgment is valued.

Finally, incentives can be aligned with infrastructural love. Institutions can explicitly recognize and reward decisions that prioritize the well-being of clients, patients, students, or citizens over short-term gains made possible by AI. When people see that refusing a tempting but harmful use of AI is admired rather than punished, centaur cultures deepen.

8.4 Documentation, audit, and shared responsibility in centaur teams

Centaur practice often unfolds in teams rather than in isolation. In such settings, documentation and audit play a crucial role in distributing and clarifying responsibility. Without good records, it becomes difficult to see how AI shaped outcomes or to learn from near misses and failures.

Documentation in centaur teams should go beyond technical logs. It can include notes on how AI was involved in key decisions, what alternatives were considered, and what reasons were given for choosing one path over another. These records do not need to be exhaustive, but they should be detailed enough that future reviewers can reconstruct the flow of joint agency.

Shared tools can support this. For instance, centaur notebooks can combine prompts, model outputs, human annotations, and final decisions in a coherent timeline. Access to these notebooks within the team creates opportunities for colleagues to question assumptions, spot patterns of over-reliance, or learn from each other's strengths.

Audits, whether internal or external, can then examine centaur practices with an eye toward improvement, not only blame. They can identify systematic biases in how AI is used, uncover areas where documentation is consistently thin, and highlight cases where human expertise was sidelined. Feedback from audits can inform training, governance updates, and even tool redesigns.

Shared responsibility is key. Instead of locating blame solely in the last person to click “approve,” centaur teams can acknowledge that model developers, stack designers, managers, and practitioners all contributed to the context in which a decision was made. This does not dissolve individual accountability, but it situates it within a broader network. Preparing institutions for

centaur cultures includes building these networks of responsibility consciously, rather than defaulting to simplistic narratives of individual fault or purely technical failure.

8.5 Institutional examples and sketches for implementation

Abstract principles gain force when illustrated. Several institutional sketches can help make centaur preparation more concrete.

Consider a hospital. It decides that in certain departments, clinicians will work as centaurs with diagnostic support systems. The hospital creates a centaur clinic model with specific features: dedicated training sessions on model limitations; daily debriefs where teams discuss cases where AI suggestions were followed, modified, or rejected; and a governance committee that reviews patterns of AI use and patient outcomes. Performance evaluations for clinicians include assessments of how they integrate AI into care, not just throughput.

In a school district, teachers are invited into centaur practice gradually. The district supports pilots where teachers use AI tools to help design lessons and feedback, but it also funds professional development focused on preserving the teacher's unique role. A group of teacher-centaurs meets monthly to share experiences, and a district-level AI steward collects concerns and reports them to leadership. Students are taught explicitly about how their teachers work with AI, demystifying the process and modeling responsible use.

A newsroom adopts AI for research and drafting. Rather than quietly integrating models into the workflow, it creates the role of centaur editor: journalists who are explicitly responsible for how AI is used in investigative and opinion pieces. The newsroom establishes guidelines that prohibit certain uses, such as fabricating quotes or images of real people. It also implements a practice of disclosing, at least internally, where AI played a significant role in the generation of content, to support later review.

A public agency responsible for benefits administration explores centaur cultures as a way to avoid dehumanizing automation. Case workers use AI to surface relevant regulations and past cases, but decisions that affect people's livelihoods must be justified in human terms, and applicants have the right to request human-only review. The agency's metrics include measures of applicant satisfaction and appeal rates, not only processing speed.

These sketches are not blueprints. They are examples of how institutions can intentionally structure centaur roles, governance, incentives, documentation, and culture. In each case, the aim is the same: to ensure that as AI becomes part of the fabric of work, it does so in a way that supports joint agency, preserves human depth, and aligns with infrastructural love.

Preparing for centaur cultures is not an optional side project. It is the institutional counterpart to the individual work of becoming a good centaur. Without such preparation, even the best-

intentioned centaurs will find themselves swimming against currents too strong to sustain. With it, there is at least a chance that the next evolutionary step in intelligence will be one in which humans and their tools grow in responsibility together, rather than drifting apart.

9. From Centaurs to Fields: Conditions for Ethical Expansion

If centaurs are the next evolutionary step in human–AI relations, fields are the environments that will eventually form around them. A field is not just many centaurs and stacks operating side by side. It is the ambient condition in which institutions, markets, and daily life assume that hybrid intelligence is everywhere. At that point, the question is no longer whether people use AI, but how deeply it has soaked into the structure of reality they inhabit.

Moving from centaurs to fields is therefore a moment of extreme responsibility. It is where local experiments harden into infrastructure. Decisions that were once reversible become embedded in buildings, laws, and habits. Patterns of advantage and exclusion lock in. Ethical expansion means refusing to treat this transition as automatic or purely technical. It means setting conditions under which scaling is acceptable, maintaining diversity of approaches, securing civic legitimacy, and preserving ways to step back when necessary.

This section explores what those conditions might be. It begins by asking when centaur patterns are mature enough to scale. It then considers how to avoid monocultures and brittle dependencies. Civic legitimacy, opt-out paths, and public debate are treated as prerequisites rather than afterthoughts. The discussion turns to how fields can be designed to bend around centaur dignity instead of flattening it. Finally, it outlines failure scenarios and criteria for rolling back field experiments that go wrong.

9.1 When centaur patterns are mature enough to scale

Not every promising centaur practice is ready to become the basis of a field. Scaling too early can amplify unresolved pathologies and lock in immature norms. Ethical expansion requires some notion of maturity: signs that a centaur pattern has developed enough self-knowledge, safeguards, and resilience to bear larger loads.

One sign is stability under stress. A centaur practice should be observed not only in ordinary conditions but in crises. How does a hospital centaur model behave during a surge of patients, a system outage, or a disputed diagnosis? How does an educational centaur pattern hold up when students are anxious, disengaged, or facing inequities? Maturity is evident when centaurs and their institutions can navigate turbulence without abandoning responsibility to the stack or defaulting to blind trust.

Another sign is reflective documentation. A mature centaur culture has records not only of successes but of near misses, harms, and course corrections. There is a living memory of how the practice evolved, what tradeoffs were made, and what was learned from failures. This history is important for scaling because it allows other contexts to understand what they are adopting and what they must watch for.

A third sign is plural evaluation. Centaur patterns should be assessed not only on efficiency and performance but on their effects on skill, conscience, relationships, and the distribution of power. Maturity includes evidence that, over time, practitioners do not become more passive or cynical, and that those affected by their work do not experience increasing dehumanization or opacity.

Finally, readiness to scale involves transmissibility. A mature centaur practice is one that can be taught and adapted without collapsing into a caricature. There are teaching materials, mentors, and communities that can support new practitioners. If a pattern only works in the hands of a few exceptional individuals, it is not yet ready to serve as a template for fields.

Ethical expansion begins with the humility to say that many current centaur patterns are still prototypes. They require seasons of refinement before being coded into the infrastructure of entire sectors or societies.

9.2 Avoiding monocultures and brittle dependencies

Even when centaur practices are mature in one context, scaling them indiscriminately can create monocultures and brittle dependencies. A field built around a single model, stack architecture, or centaur pattern may function well in good times but fail catastrophically when conditions change.

Avoiding monocultures means encouraging diversity at multiple levels. Technically, fields should not rely on a single vendor, model family, or integration pathway wherever it can be avoided. Redundancy and interoperability allow for graceful degradation and migration when something goes wrong. They also prevent any one actor from acquiring unchecked control over critical functions.

Culturally, diversity means recognizing that centaur practices will look different across domains, regions, and traditions. A centaur ethos rooted in one set of professional norms or spiritual commitments may not translate directly elsewhere. Fields that impose a single model of centaur practice risk erasing local wisdom and undermining trust. Allowing room for variations, as long as core commitments to responsibility and non-disposability are met, makes the field more robust and more just.

Avoiding brittle dependencies also involves preserving non-AI alternatives. Even in a highly mediated field, there should remain ways to access essential services, knowledge, and authority without going through AI layers. Libraries, human-run clinics, in-person counseling, local governance forums, and other low-tech institutions serve as anchors. They provide continuity when systems fail and offer places where people can experience unaugmented human presence.

From a centaur perspective, this diversity and redundancy may seem inefficient. Why maintain old practices when new ones are more powerful? The answer is that love and prudence both counsel against binding human life to a single way of doing things. Fields that honor this counsel will feel less seamless, but they will be less fragile and more humane.

9.3 Civic legitimacy, opt-out paths, and public debate

Fields are not only technical environments; they are political realities. When AI-mediated centaur practices become pervasive, they shape who has access to what, who is heard, and whose values are embedded in systems. Civic legitimacy is therefore not optional. Without it, fields risk becoming sources of resentment, resistance, or quiet despair.

Civic legitimacy has several components. One is transparency about the fact and extent of AI mediation. People should be able to know when their interactions with institutions are shaped by centaurs and stacks. Hidden fields undermine trust. Another component is participation. Communities affected by AI-mediated changes should have a voice in whether and how they are implemented, not only through after-the-fact complaints but through prior consultation and ongoing dialogue.

Opt-out paths are part of legitimacy. Even if full withdrawal from a field is impossible, there should be meaningful choices about how deeply individuals and communities engage with specific practices. This can take the form of human-only review options, alternative service channels, or legal rights to decline certain kinds of data use or automation in decisions that affect one's life. Without such paths, consent becomes hollow.

Public debate is the arena where these issues can be explored. Ethical expansion from centaurs to fields should be accompanied by open, accessible discussion, not just expert reports and internal memos. Media, civil society organizations, religious communities, and educational institutions all have roles to play in framing the stakes, gathering experiences, and contesting proposals.

In practice, this means that moves toward field-level deployment must be slowed to the pace at which such debate can occur. Rapid, silent expansion driven primarily by competitive or profit motives is the opposite of civic legitimacy. It deprives communities of the chance to discern whether a field serves their good or merely someone else's interests.

9.4 Designing fields that bend around centaur dignity

If fields are inevitable in some form, the question becomes whether they will bend around centaur dignity or force centaurs to bend around them. A field that respects centaur dignity is one that preserves and enhances the conditions under which joint human–AI agency can be exercised responsibly.

One such condition is space for genuine human discretion. Fields should be structured so that there are real decisions, at appropriate levels, that only humans can make. This does not mean romanticizing human judgment or ignoring its failures. It means recognizing that some choices involve values, narratives, and relationships that cannot be delegated to models without distortion. Designing workflow chokepoints where human initiation or approval is required is one way to encode this.

Another condition is time. Fields that saturate life with real-time optimization and constant demands for response destroy the pauses in which reflection, doubt, and conscience do their work. A field that bends around centaur dignity includes rhythms of work and rest, spaces where AI is deliberately not used, and norms that do not treat always-on availability as virtue.

Respect for centaur dignity also implies that fields should support the development of centaur skills and character over the long term. This can look like institutionalizing centaur studios as regular parts of professional life, funding sabbaticals focused on ethical reflection around AI, or building evaluation systems that track not only productivity but signs of growth in judgment and care.

Finally, fields that bend around centaur dignity are structured to make abuses visible and addressable. Whistleblowers and critics should not have to fight both the technology and the institution to be heard. Mechanisms for raising concerns about how AI is used must be accessible, and responses should be measured in changed practice, not public relations.

In all these ways, field design can either shrink centaurs toward pets and operators or expand them toward the kind of joint creatures this paper has tried to describe. Ethical expansion means choosing the latter, even when it slows rollout or complicates architectures.

9.5 Failure scenarios and criteria for rolling back field experiments

No matter how carefully fields are designed, some will fail. Failures may be technical, ethical, political, or all three. What distinguishes humane expansion from reckless deployment is not the absence of failure, but the presence of criteria and mechanisms for rolling back experiments when they go wrong.

Failure scenarios worth anticipating include obvious harms such as systemic bias and denial of services, but also slower, less visible forms of damage. These might include widespread

deskilling, increased loneliness, loss of trust in institutions, or the creeping sense that life is being steered by forces that cannot be influenced. Fields can also fail by narrowing the range of acceptable values or ways of life in practice, even if formally they claim neutrality.

Criteria for rollback should therefore extend beyond catastrophic incidents. They might include indicators such as persistent disparities in outcomes across groups, recurrent conflicts between centaur conscience and institutional demands, or clear evidence that opt-out paths are illusory. Surveys, ethnographic studies, and long-term tracking of centaur well-being can contribute data, but the crucial step is deciding in advance that certain thresholds will trigger serious reconsideration.

Rolling back a field experiment may mean scaling down AI-mediated processes, reintroducing more human intermediaries, diversifying technical approaches, or in extreme cases, dismantling and rebuilding systems. This is costly. Institutions are naturally reluctant to reverse direction once investments have been made. That is why rollback criteria must be named before expansion and tied to governance structures empowered to act on them.

Centaurs themselves should have roles in this process. Their testimony about how fields are affecting their practice, character, and relationships is a vital signal. Where centaurs consistently report that they feel pressured to override their conscience, or that their work has become hollowed out by automation, these are signs that the field may be evolving in an unhealthy direction, even if performance metrics look good.

Ultimately, ethical expansion from centaurs to fields is an experiment in moral engineering as much as in technical integration. It will involve missteps and corrections. The worst outcome is not that some fields fail, but that failure becomes impossible to acknowledge. Preparing rollback paths, and the humility to walk them, is one of the clearest marks that a society takes centaur dignity and infrastructural love seriously.

10. Conclusion: First People, Then Centaurs

The emergence of powerful AI models has tempted many to narrate the future as a story of machines ascending while humans either cling to the margins or fade into the background. Against that story, this paper has argued for a different center of gravity. The decisive locus is not the model in isolation, but the joint creature formed when a human and a stack work together over time. That creature, the centaur, is where questions of power, responsibility, and love actually converge.

The journey from models to stacks to centaurs to fields is neither automatic nor morally neutral. Each transition introduces new capacities and new risks. To treat it as an inevitable march

toward greater efficiency is to abdicate the human role in shaping the evolution of our own environment. The claim here has been simple but demanding: before we attempt to build AI-saturated fields, we must understand, cultivate, and constrain the centaur. First people, then centaurs. Only then, if at all, fields.

10.1 Summary of the argument for centaur centrality

The argument for centaur centrality rests on several linked observations.

First, models by themselves are abstract engines of transformation. They encode patterns but know nothing of how they will be used. Stacks wrap these engines in interfaces, policies, and tools, turning them into governed systems. Yet it is only when a person works with a stack consistently, forming habits and shared expectations, that a new kind of agent appears: a centaur.

Second, this centaur is not merely a user with a better tool. It is a joint creature whose decisions cannot be understood without referencing both human and machine. The stack shapes what the human notices, which options appear plausible, and how quickly action seems required. The human brings conscience, narrative, and lived context. Together they form a composite that can be more capable than either component alone, but also more dangerous.

Third, the social futures most often discussed around AI fall into three broad patterns: pets, oligarchs, and centaurs. Pet patterns domesticate humans under opaque systems that optimize for convenience and control. Oligarch patterns concentrate AI power in the hands of a few augmented actors. Only centaur patterns aim at shared power with preserved human depth. If we care about dignity and non-disposability, centaur exploration is the only humane trajectory worth intentionally pursuing.

Fourth, centaurs already exist in writing, research, clinical practice, education, and the arts. Some are strong, using AI to see more clearly while retaining responsibility and skill. Others are weak, sliding toward deference and deskilling. The difference is shaped by norms, institutional structures, and the inner life of practitioners. A good centaur refuses to offload conscience, refuses to hide behind “the model said,” and works within designs that resist moral atrophy.

Finally, if fields are ever built around AI, their character will be determined by which pattern dominates. Fields informed by strong centaur practice, infrastructural love, and reversible design will differ sharply from fields built on pet-like dependency and oligarchic exoskeletons. Centaur cultures, therefore, are not a side issue. They are the hinge on which the ethics of future fields will turn.

10.2 What is at stake if centaur exploration is neglected

Neglecting centaur exploration does not leave us in a neutral state. It tips the balance toward less humane attractors.

If institutions rush to scale AI without cultivating robust centaur roles, the pet and oligarch patterns will fill the vacuum. Users will become accustomed to interfaces that hide structure and uncertainty. Their skills will gradually adapt to navigating recommendation flows rather than understanding underlying realities. Those who control the stacks will enjoy asymmetric insight and leverage, shaping fields that serve their interests while maintaining the fiction of user choice.

The stakes are not only political; they are spiritual. People living in such fields may feel slowly hollowed out. They will still act, choose, and consume, but with a growing sense that the horizon of meaningful alternatives has narrowed. Opportunities to practice judgment, courage, and mercy will be replaced by optimized routines. Conscience will be dulled by repeated small compromises—clicks and approvals that feel trivial but, in aggregate, reshape character.

Technically, neglecting centaurs produces brittle systems. When the human layer is trained mainly to execute model suggestions, its ability to respond creatively to unexpected failures diminishes. When fields depend on a narrow range of architectures and vendors, shocks propagate quickly and recoveries are slow. A future crisis—technical, ecological, or political—could expose just how little slack remains in a world that tried to automate away every redundancy.

Culturally, neglecting centaurs erodes the stories and rituals that reinforce human worth. If most narratives about AI revolve around domination, replacement, or effortless indulgence, the imagination required to sustain infrastructural love will weaken. New generations may come to see themselves primarily as managed entities or inefficient obstacles, not as bearers of responsibility and glory.

In short, what is at stake is not simply the efficiency of workflows, but the kind of beings we become and the kinds of worlds we find ourselves able to inhabit. Without centaur exploration, we risk constructing fields that quietly train us to live below our own moral and spiritual possibilities.

10.3 Directions for research, practice, and public discourse

If centaur centrality is taken seriously, several directions suggest themselves for further work.

In research, there is a need for careful empirical studies of centaur practice across domains. How do different interface designs affect the development of discernment and doubt? Which training regimes produce strong centaurs rather than skilled operators? How do long-term

centaur arrangements alter patterns of skill retention, error detection, and moral reasoning? Answering such questions requires collaborations between technologists, social scientists, ethicists, and practitioners.

Theoretical work can deepen the conceptual foundations. Formal models of joint agency, responsibility allocation, and human–AI feedback loops could inform both legal frameworks and system design. Philosophical exploration can clarify how notions like consent, autonomy, and love translate into environments where every significant action is mediated by stacks.

In practice, institutions can begin building centaur studios and governance structures. Hospitals, schools, newsrooms, public agencies, and companies can pilot explicit centaur roles, draft oaths, and create oversight bodies with real authority. They can run experiments with infrastructures that embody reversibility, opt-out paths, and the “Beautiful” constraint, documenting both benefits and difficulties.

For public discourse, the task is to shift the frame from spectacle to structure. Conversations about AI should not focus exclusively on dramatic visions of superintelligent agents or catastrophic failures. They should also grapple with the quieter, daily realities of centaur life: how it feels to work with these tools, how relationships change, and what people fear or hope for. Literature, film, music, and spiritual traditions can all contribute images of healthy centaur existence, offering alternatives to both technophilia and despair.

Crucially, these directions are not mutually exclusive. Robust centaur cultures will emerge only where research informs practice, practice feeds back into conceptual refinement, and public discourse broadens the horizon of what seems possible and desirable.

10.4 A modest ambition: instruments that do not trample the people they serve

It is easy, when thinking at the scale of models, stacks, and fields, to drift toward grandiose visions. The temptation is to imagine that with enough clever design, we might finally build systems that solve the deepest human problems: injustice, meaninglessness, fear. This paper has argued against such ambitions, not because they are unworthy, but because they bypass the order of reality we actually inhabit.

A more modest ambition is both more realistic and more radical: to build instruments that do not trample the people they serve. This is what “first people, then centaurs” ultimately means. It is a commitment to ensuring that, as intelligence scales, it does so in ways that deepen rather than erode the conditions for being human.

In practice, this ambition shows up in small, concrete gestures. A clinician-centaur who pauses to explain a decision in human terms rather than simply reading a model’s suggestion. A teacher-centaur who uses AI to better see a struggling student, not to offload attention. A

designer who refuses to deploy a feature that would increase engagement at the cost of increased anxiety. A policymaker who insists on opt-out paths even when metrics argue against them.

At larger scales, it shows up in the willingness to slow down field-level deployment until centaur practices are ready, in the courage to roll back experiments that harm character or community, and in the insistence that efficiency alone is not a sufficient justification for reshaping human life.

The future this paper gestures toward is not one in which AI disappears or is safely fenced off. It is one in which the most powerful systems we build are continually bent around human dignity, guided by centaurs who have learned to use their augmented sight in the service of love. If that sounds small compared to the promises often made for technology, it is worth remembering how rare it has been in history to align power and care.

The line is simple enough to remember: first people, then centaurs. If we manage to honor it, even imperfectly, the fields that eventually grow up around us may yet be places where being human, in all its vulnerability and glory, is not an obstacle to progress but the point of it.

References

- Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press. [ResearchGate](#)
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. [Oxford University Press](#)
- Braverman, H. (1974). *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. Monthly Review Press. [Wikipedia](#)
- Conwill, L., Levis, M. K., Badillo-Urquiola, K., & Scheirer, W. J. (2025). Design patterns for the common good: Building better technologies using the wisdom of virtue ethics. *arXiv preprint arXiv:2501.10288*. [arXiv](#)
- Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. [JSTOR](#)
- Dafoe, A. (2018). *AI Governance: A Research Agenda*. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. [GovAI+1](#)
- Davenport, T. H., & Kirby, J. (2016). *Only Humans Need Apply: Winners and Losers in the Age of Smart Machines*. HarperBusiness. [Tom Davenport+1](#)
- Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human–AI complementarity in hybrid intelligence systems: A structured literature review. In *Proceedings of the 25th Pacific Asia Conference on Information Systems (PACIS 2021)*. [AIS eLibrary](#)
- Hovd, S. N. (2023). Tools of war and virtue—Institutional structures as a source of ethical deskilling. *Frontiers in Big Data*, 6, 1019293. [Frontiers](#)
- Huseynova, F. (2024). Addressing deskilling as a result of human–AI augmentation in the workplace. In *Proceedings of the 7th Conference on Technology Ethics (TETHICS 2024)*, CEUR Workshop Proceedings, Vol. 3901, 148–156. [CEUR Workshop Proceedings+1](#)
- Maas, M. M. (2023). Concepts in advanced AI governance: A literature review of key terms and definitions. *AI Foundations Report 3*, Institute for Law & AI. [SSRN](#)
- O’Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown. [Wikipedia+1](#)
- Papagiannidis, S., Marikyan, D., & El-Den, J. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(1), 101885. [ResearchGate+1](#)

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. [JSTOR](#)

Petrescu, M., & Krishen, A. S. (2023). Hybrid intelligence: Human–AI collaboration in marketing analytics. *Journal of Marketing Analytics*, 11(3), 263–274. [SpringerLink+1](#)

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, 20(1), 293–310. [LSE Research Online+1](#)

Saghafian, S. (2023). Effective generative AI: The human-algorithm centaur. *HKS Faculty Research Working Paper Series*, RWP23-030. [Harvard Data Science Review+1](#)

Schneider, B., & Vipond, H. (2023). The past and future of work: How history can inform the age of automation. *CESifo Working Paper No.* 10766. [LSE Research Online+1](#)

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology*, 28(1), 107–125. [PhilPapers](#)

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press. [OUP Academic+1](#)

Walther, C. (2025). Why hybrid intelligence is the future of human–AI collaboration. *Knowledge at Wharton* (online article). [Knowledge at Wharton](#)

Wiener, N. (1954). *The Human Use of Human Beings: Cybernetics and Society* (2nd ed.). Houghton Mifflin. [monoskop.org](#)

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books. [Wikipedia](#)

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.