

**PHRONEXUSworks LLC™**

**Experience | *Wisdom* | Results**

www.phronexusworks.com

# “Conscience-ness” in Human–AI Teams

*How do we keep ethical judgment effective as people and AI act together?*

**John A. Poirier**

*Developed in dialogue with ChatGPT*

October 21, 2026

A *PHRONEXUSworks*™ LLC publication.

© 2026 John A. Poirier. All rights reserved.

A vigorous debate over AI consciousness and its implications is ongoing. Let’s examine whether that is the essential question for the work that human-AI ecosystems are doing together. The inquiry itself raises others. Is the immediate issue consciousness, or the implications of the behaviors and actions a human–AI combination undertakes? What would knowing that AI is conscious change about the decisions we face today? If AI is conscious, does its ethical foundation align with human ethical intent?

Consciousness raises questions about experience, awareness, and how understanding arises. Those questions remain open to exploration. Meanwhile, human–AI combinations already produce consequences for people who may have little involvement in the decisions affecting them. We need a way to examine that *practical* reality, including how the combination develops its understanding of what it should do.

We establish rules, provide instruction, and reward behaviors. Yet we still have to examine what happens when those arrangements encounter a situation nobody anticipated. Possessing instructions about acceptable behavior leaves us with the continuing challenge of interpreting and applying them as purposes, relationships, and conditions change. Can ethical behavior be programmed into AI? Equally important, what is the human equivalent that drives how AI even understands ethical dimensions? These questions become increasingly important as the scale of interaction changes and problems being addressed become harder.

We can explore this as “conscience-ness”: the continuing practice of bringing ethical judgment into how we define hard problems, choose and carry out actions, and assess their anticipated and experienced consequences for those affected.

Conscience-ness is offered as a starting point for discussion. It distinguishes a practical concern from the metaphysical question of consciousness. We can examine ethical implications through observable decisions, behavior, and consequences while leaving questions about subjective experience open. The focus here is the co-evolving human–AI socio-technical ecosystem through which the work occurs.

The focus in this inquiry is on what helps ethical concerns influence that system’s direction. Existing work on human responsibility and AI accountability, including oversight, provides important foundations. Driving the focus is *reciprocal challenge* as a way to improve alignment of human-AI behaviors. The idea is that reciprocal challenge offers a mechanism for humans to monitor and adapt AI performance and for AI to feed back to humans regarding issues that may not have been considered. Specifically, it seeks a way for AI to be “programmed” or prompted to provide analysis across roles and scales to anticipate ethical issues as behaviors propagate and challenge the human to confirm that the potential risks and results are acceptable. The goal is to ensure purpose and values across scales, and the practical capacity to exercise responsibility, are clear. *PHRONEXUSworks*™, an operational architecture for navigating hard problems, offers a way to organize this inquiry.

## Whose purposes and values direct the work

A basic proposition is that purpose gives a team direction. Values influence what it considers worthwhile, acceptable, or unacceptable. Both need examination. A team can faithfully pursue a clear purpose while creating consequences that others have good reason to challenge.

The proposition raises questions. Whose purposes and values shape the work? Who experiences its consequences? Who has standing and opportunity to question them? These questions become more consequential as AI contributes to how we interpret situations, develop alternatives, and execute decisions.

The human–AI combination can operate as a team within a wider team of teams. An individual or group working with AI may contribute to a project or endeavor, which works within an organization accountable to institutions and society. Human roles become differentiated to allocate work. AI can also assume multiple roles such as collaborator, developer, researcher, reviewer, and others. It executes these roles while seeming to be a single entity. The human-AI team then functions as a co-evolving ecosystem working towards achieving an outcome. That ecosystem’s roles need defined expectations and the ability to adapt as circumstances change.

Across these relationships, people interpret purposes through different responsibilities, pressures, and available information. An organization may value careful judgment while rewarding rapid completion. A manager’s request for a preliminary recommendation may become someone else’s instruction to produce a finished answer. Qualifications can disappear while the deadline remains visible. AI may offer rapid access to and processing of voluminous information. How tasks are prompted influences its contribution, alongside training, system instructions, available context, and permissions. Prompting becomes an important interface through which the human–AI team establishes what to examine and act upon.

Ethical behavior therefore depends on more than the values formally stated. It also depends on how those values are interpreted, resourced, and expressed through actual decisions. People can make the connection explicit by identifying affected groups, articulating competing obligations, and specifying what must be examined before action. AI can then be tasked to expose possible implications across those relationships, including effects beyond the immediate team. Conscience-ness gives us a question to keep asking: what is this arrangement encouraging us to do, and with what consequences for others?

© 2026 John A. Poirier | PHRONEXUSworks™ LLC

## Responsibility while someone else acts

Consider a ship running aground while the captain is asleep. This analogy illustrates the relationship among responsibility, authority to act, and execution. The officer of the deck, conning officer, and helmsman may have been directing and executing the immediate actions. The captain’s responsibility continues, but authority to act has been delegated. The practical concern is how responsibility remains effective when others exercise delegated authority. Similarly, tasking AI to find, analyze, and parse information does not diminish the responsibility of the human tasker for outcomes.

A starting principle is that people and human institutions remain the ultimately responsible agents in human–AI systems. AI can receive bounded authority to act. People remain responsible for establishing purposes, authorizing activity, supporting oversight, and responding to consequences.

| Distinction | Question for the team | What makes it effective |
|---|---|---|
| Responsibility | Who must ensure appropriate action and respond to consequences? | Identifiable human obligations supported by information and capacity. |
| Authority to act | What may each participant decide or do? | Defined permissions, limits, and routes for approval or escalation. |
| Execution | Who or what carries out the work? | Actions within authorized bounds, with opportunities to examine and intervene. |

Those distinctions become harder to maintain as participation expands. An operator may assume the organization has evaluated the system; the organization may assume an attentive operator will catch problems. The developer may have anticipated a use quite different from the one now occurring. Responsibility becomes clouded in the gaps among those assumptions.

It can also become concentrated on someone with limited influence. Madeleine Clare Elish’s research on “moral crumple zones” examines how a human actor can receive blame despite having limited control over an automated system.[1] That raises a practical test: does the person assigned responsibility have the information, authority, capacity, and opportunity to exercise it?

The same question applies at organizational and institutional scales. Assigning responsibility establishes an obligation. Making it effective requires arrangements through which the obligation can influence what happens.

## Making reciprocal challenge part of the team

Human–AI teamwork introduces an opportunity to examine the direction of the work from more than one contribution. Humans can challenge AI outputs, assumptions, and proposed actions. AI can be configured and prompted to identify conflicts, omissions, or possible consequences in human requests. Mutual monitoring and feedback can then allow the strengths of each participant to be used and limitations to be mitigated. This expands the notion of enabling reciprocal challenge.

© 2026 John A. Poirier | PHRONEXUSworks™ LLC

Reciprocity concerns the contributions to inquiry. Human responsibility remains explicit. A useful AI warning does not establish moral agency, and an apparently careful explanation still needs examination. Likewise, human approval needs support from understanding and the ability to intervene.

Imagine a hypothetical team developing an AI curriculum under a tight deadline. Its purpose includes strengthening practitioners’ judgment, but development time limits the opportunity to create and examine meaningful exercises. AI helps produce course materials and assessments based on human direction. The exercises provide insufficient evidence that participants can demonstrate the intended capability, but the AI produces them as tasked. Despite the constraints that limit the product, the team then proposes describing graduates as ready to apply AI in consequential work. Expectations are set and the organization and institution treat the product as meeting the goals.

An AI review then flags that the exercises mainly demonstrate performance with assistance. Humans challenge that concern: effective use of assistance is part of the intended capability. The exchange exposes a further question. What evidence would show that participants can recognize when assistance is inappropriate, question its output, and explain their choices? The team revises its assessment and qualifies the readiness claim before running a pilot. Expectations then have to be managed.

Tasking AI to identify potential consequences from the outset could have exposed these limitations before readiness expectations were established. Reciprocal challenge could then have informed both problem definition and the proposed solution. The ethical concern becomes concrete when others may rely on a claim that the available evidence does not adequately support.

We cannot assume that this exchange will happen reliably. The workflow must make concerns visible and provide access to someone able to respond. A warning that nobody can respond to provides little protection. Equally, an invitation to challenge carries little weight if the schedule, incentives, or permissions make reconsideration impractical.

Challenge also needs an ethical reference point. A team might question its assumptions thoroughly and become more effective at pursuing a harmful purpose. We therefore need to connect challenge with affected people, competing values, evidence, and legitimate authority. Who can question the purpose itself, and what makes that question consequential?

With that in mind, we can consider reciprocity as an aspect of human-AI relationship engineering: deliberately shaping how participants contribute, question, authorize, and learn together. Used in this practical sense, it asks how the relationship should function. Prompt engineering can help express those expectations in a particular task, while the workflow gives people the means to act on what the exchange reveals.

In the example, a request could ask AI to examine a proposed action for implications affecting an individual, the project team, the organization, and people beyond it; identify conflicts with the stated values; and distinguish supported concerns from possibilities needing investigation. It could also specify which uncertainties require clarification before execution. Humans would then examine those concerns, seek viewpoints the exchange lacks, and determine the response. However, an instruction to identify ethical implications supports analysis; it cannot guarantee that every consequential issue will be found.

© 2026 John A. Poirier | PHRONEXUSworks™ LLC

The performance proposition is worth examining. Reciprocal challenge may help a team expose omissions, calibrate confidence, and avoid committing resources to a poorly framed or constrained task. It also consumes attention and can generate weak objections. The useful test is whether the exchange improves the quality and consequences of the team’s work enough to justify that effort. Agreement and repeated challenge each need judgment about when to proceed.

## Organizing understanding and solution making

This is where *PHRONEXUSworks™* can help organize the effort. Its contribution is to relate questions that are often considered separately, so that understanding, capability, action, and learning inform one another.

It starts by distinguishing problem-solving from solution-making. Problem-solving can address definable portions of a situation, determine what relationships and stakeholders are involved, and define the range of actions needed to address a problem. Solution-making considers how those contributions work together within changing conditions and relationships to develop the relevant options for action. In the curriculum example, problem-solving helps define and address particular learning and assessment challenges. Solution-making connects those contributions into an approach for developing the intended capability, implementing it, and adapting as consequences emerge.

This lets us look more precisely at the Situation, Problem, and Decision Spaces. These offer a framework for organizing understanding. We examine the setting and affected participants, frame what needs attention, and identify the choices and commitments required. Movement among the spaces helps us revisit an assumption when something consequential comes into view. Decision-making support develops that understanding; decision support helps examine the available choices.

To keep this manageable, we can look at a small set of architectural connections and the questions they bring into view.

| Architectural connection | Practitioner question |
|---|---|
| Problem-solving and solution-making | How does completing this task contribute to addressing the wider situation and its consequences? |
| Situation, Problem, and Decision Spaces | What do we need to understand, frame, and decide as the picture develops? |
| Learning, Production, and Resource Management cycles | How do learning, delivery, and resource commitments influence one another? |
| Seven assessment tests | Is the approach useful, efficient, effective, fair, legitimate, adaptable, and sustainable? |

Taken together, these relationships connect Mindset, Methods, and Mechanisms: a way to think, a way to do, and a way to make. The cycles keep learning, production, and resource management connected, so that what we discover can change what we produce and how we support it. A deadline, for example, affects what can be produced and what can be examined, with consequences for the capability the team develops.

© 2026 John A. Poirier | PHRONEXUSworks™ LLC

Within that continuing interaction, applying the seven tests provides a mechanism to examine conscience-ness in practice. They help us assess whether ethical intentions find expression in the arrangement and its consequences. Fairness and Legitimacy may expose concerns that favorable performance and cost figures leave unresolved, while Adaptability and Sustainability extend the inquiry through change and time. Their use supports continuing examination as experience develops.

That examination also depends on what supports our understanding. Consilience offers a useful connection: different lines of evidence from different viewpoints can converge on a common explanation. In Whewell’s account, an explanation gains particular support when it also accounts for kinds of findings beyond those used to develop it.[4] We can explore that idea without expecting every viewpoint to yield the same conclusion.

For the curriculum team, observations of participants’ reasoning, performance in unfamiliar situations, and feedback from subsequent practice might support a judgment about readiness. We would need to examine their quality, independence, and relevance. Several AI reviews based on the same material could repeat an assumption while creating an appearance of corroboration. Human experience and evidence from affected participants can challenge what the team and its AI have come to accept. Reciprocal challenge can help the team identify which questions each participant should examine and where additional evidence or viewpoints are needed.

Consilience can strengthen the basis for a judgment; sufficiency to act also depends on uncertainty, consequences, and opportunities for correction. Another consideration is how agreement enables or inhibits effective execution. Agreement about an explanation still leaves ethical choices about action to examine. Bringing those considerations together helps us ask what warrants proceeding and what would cause us to reconsider. This connects the evidence informing a decision with the conscience-ness through which we examine its implications.

## Influencing what the arrangement becomes

The same connection between ethical intention and practical behavior applies to how AI is developed and used. Constitutional AI research describes one approach using explicit principles and AI feedback during training.[2] Through deployment, those influences encounter purposes, permissions, and human interpretations that require continuing attention. NIST’s AI Risk Management Framework similarly relates trustworthiness to organizational behavior, design decisions, and human oversight.[3]

As the team identifies implications, it can respond by avoiding an activity, mitigating its consequences or likelihood, accepting a bounded exposure under appropriate authority, or transferring specified obligations. Opportunities can also be pursued. Choosing a response brings us back to purpose and values: who benefits, who bears the consequences, and what responsibility continues?

Those choices, in turn, influence subsequent work. As the experience of the human-AI team accumulates, people change how they delegate, check, and interpret, while teams adjust permissions and expectations. Organizations revise what they reward and require. AI contributions are shaped by instructions, selected context, and changes to the systems themselves. Co-evolution includes these changes in the working arrangement; it does not depend on every exchange retraining a model.

© 2026 John A. Poirier | PHRONEXUSworks™ LLC

That is why ethical judgment needs continuing expression. A boundary appropriate for an early experiment may need revision when its use expands. An apparently successful practice may increase reliance while weakening the capability needed to recognize its limits. Learning from consequences must be able to change future commitments.

Conscience-ness brings these concerns together as a practitioner question. Human responsibility persists through delegated action, and its exercise depends on what people can understand, authorize, question, and change. Reciprocal challenge is one team dynamic that may help keep ethical concerns active across those relationships.

This proposal is open to examination and exploration. What would make this approach useful in your work? Where does it leave an important relationship outside the aperture?

Most directly, when a concern arises in your human–AI team, what allows it to change the direction of the work—and who has the responsibility and practical means to make that happen?

These ideas build on *Figuring Out Figuring Out* and *Anticipating Co-Evolution*. The *PHRONEXUSworks™* website is a community resource introducing the architecture and linking to the books: https://phronexusworks.com. The invitation is to find something useful to question, adapt, or apply, and contribute to how the thinking develops.

## Sources informing this exploration

[1] Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society, 5, 40–60. https://doi.org/10.17351/ests2019.260

[2] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. Research preprint. https://doi.org/10.48550/arXiv.2212.08073

[3] NIST (2023). Artificial Intelligence Risk Management Framework, AI RMF 1.0, section 3, AI Risks and Trustworthiness. https://doi.org/10.6028/NIST.AI.100-1

[4] Snyder, L. J. William Whewell, section 3, Philosophy of Science Confirmation. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/whewell/ (accessed October 20, 2026).

## Preparation note

This article is part of continued experimentation with human-AI interactions. This article was developed through iterative collaboration with ChatGPT, using the author’s concepts, questions, and direction. The author modified, directed, and revised direction through multiple drafts. AI assisted with research discovery, structure, drafting, and revision. The curriculum example is hypothetical and illustrates the argument; it does not report an empirical evaluation. The author retains responsibility for the content.

© 2026 John A. Poirier | PHRONEXUSworks™ LLC