Beyond Capabilities: A Framework for Mission-First AI Adoption in Identity-Critical Contexts

Frank Brockmann
Independent Researcher, Folsom, CA, USA
frank@centerpointcorp.com

January 2026

Abstract

Organizations increasingly adopt artificial intelligence based on technical capabilities rather than mission alignment, risking what organizational theorists call "value drift": the gradual erosion of core purpose in favor of operational efficiency. This risk is amplified in identity-critical contexts, where organizational activities shape individual or collective identity formation. This paper proposes a Mission-First Adoption Framework that inverts the typical adoption sequence: articulate a mission core, define explicit boundaries on prohibited applications, and only then evaluate AI capabilities within a "permissible sandbox." Two design cases—a Hawaiian-medium educational assessment program (KĀʻEO) and a nonprofit eating disorder support organization (F.E.A.S.T.)—illustrate how mission-driven governance can enable responsible AI integration. The framework is positioned as an operationalization of Value Sensitive Design principles for organizational AI adoption, offering a methodology for translating abstract values into concrete governance constraints.

Keywords: AI governance, mission-driven design, Indigenous data sovereignty, Value Sensitive Design, identity-critical contexts, organizational AI adoption

1 Introduction

When organizations adopt AI today, they typically follow a capability-first pattern. They encounter a conversational interface built on large language models (LLMs), such as ChatGPT or Claude, and, impressed by its speed, fluency, and apparent reasoning, immediately ask: *How can we use this?* Governance typically comes later, treated as a set of guardrails added to a system already in deployment.

Recent surveys confirm this pattern is widespread. A 2024 McKinsey Global Survey found that 65 percent of organizations regularly use generative AI, yet only 18 percent have an enterprise-wide council or board with authority to make decisions involving responsible AI governance (Singla et al., 2024). The gap is stark in mission-driven organizations: a Stanford Human-Centered Artificial Intelligence and Project Evident survey found that 66 percent of nonprofit respondents use some type of AI, yet 78 percent have no organizational policy guiding that use (Di Troia et al., 2024). An ISACA survey found that only 10 percent of organizations have a formal, comprehensive policy for generative AI, while 41 percent of respondents believe insufficient attention is given to ethical standards in AI implementation (ISACA, 2023). The pattern is clear: adoption outpaces governance by a wide margin.

For many tasks (code editing, email drafting, meeting summaries), this approach is practical. The cost of errors is low, iteration is fast, and the worst outcome is usually minor embarrassment or wasted time. But for what this paper terms *identity-critical* work, where outcomes affect how people understand themselves, their culture, or

their wellbeing, this sequence introduces substantial risk.

This paper defines **identity-critical contexts** as domains where organizational activities directly shape individual or collective identity formation: Indigenous language preservation, mental health support, cultural education, religious guidance, and similar areas where stakes extend beyond task completion to the formation of self-understanding. In these contexts, the efficiency of a tool may directly conflict with the mission of the organization deploying it.

This concern is not merely theoretical. Organizational scholars have documented how hybrid organizations (those pursuing both social missions and operational sustainability) face systematic pressures that cause financial considerations to override social purpose (Ebrahim et al., 2014). Bruder (2025) extends this analysis, distinguishing "practice drift" (changes in operational routines) from full "mission drift" (fundamental purpose erosion), noting that drift processes are often "unintended, unnoticed, and unwanted." This paper uses "value drift" as an umbrella term encompassing both phenomena, emphasizing the erosion of core values that defines the risk. AI adoption may accelerate these dynamics: when a capability-first evaluation determines that a chatbot can "scale support," organizations may adopt it without recognizing that the very efficiency it provides undermines the relational mechanisms that make support effective.

Smaller, resource-constrained organizations may be particularly vulnerable. The efficiency gains that make AI attractive also create pressure to deploy systems without adequate governance structures. An understaffed nonprofit struggling to respond to community needs faces genuine tension between operational sustainability and careful technology evaluation. Yet these organizations, often serving vulnerable populations, may have the most to lose from value drift.

This paper argues for inverting the adoption sequence. It proposes a **Mission-First Adoption Framework** (hereafter, "Mission-First") where organizations define their boundaries *before* evaluating any technology. By articulating what AI systems are explicitly *forbidden* from doing, even when technically capable, organizations can innovate more safely within clearly defined constraints. Drawing on design cases from an Indigenous assessment program and an eating disorder support organization, the paper illustrates that mission-first governance functions not merely as a safety measure but as a design tool that enables responsible innovation.

Contributions. This paper contributes:

1. A mission-first adoption sequence that foregrounds prohibited-use "negative space" as a governance design move;
2. An identity-critical rubric to scope when the framework applies; and
3. Two design cases showing how organizational values were translated into enforceable AI governance constraints in practice.

2 Theoretical Foundations

2.1 The Capability-First Trap

The "Capability-First Trap" occurs when technology defines organizational boundaries rather than mission defining them. In a capability-first model, the guiding question is: *What can this system do for us?* This framing inverts sound organizational design. When technology capabilities determine what is possible rather than mission determining what is appropriate, organizations risk **value drift**, the gradual erosion of core purpose in favor of operational efficiency.

Consider a mental health organization whose mission depends on authentic human connection between counselors and clients. Current large language models can generate text that reads as deeply empathetic; they can respond to emotional disclosures with apparent warmth, validation, and understanding. A capability-first evaluation might conclude that an LLM-based chatbot could "scale support" by providing 24/7 availability and consistent responses. The technology is certainly *capable* of this application.

But if the organization's mechanism of change depends on the therapeutic relationship (on what

the psychotherapy literature terms "working alliance" (Bordin, 1979; Horvath and Symonds, 1991)), then simulated empathy may represent a category error. The LLM produces text that *resembles* empathic response without the relational foundation that research suggests makes empathy therapeutically effective. Even if the chatbot satisfies operational metrics (response volume, availability, user satisfaction scores), it may hollow out the mission's core mechanism. The capability becomes a liability precisely because it is so convincing.

This risk intensifies as models improve. Early chatbots were obviously mechanical; current systems feel conversational, even wise. This "smoothing effect" masks the underlying architecture. Risk management guidance for generative AI emphasizes that improved capability and naturalistic interaction can exacerbate harms, including misleading outputs and unsafe recommendations, making governance decisions increasingly consequential as systems become more persuasive (National Institute of Standards and Technology, 2024).

2.2 Value Sensitive Design and Its Limitations

The Mission-First Framework builds on Value Sensitive Design (VSD), the tripartite methodology developed by Friedman and Hendry (2019) that integrates conceptual, empirical, and technical investigations to embed human values in technology design. VSD asks not only "whether we CAN develop a technology" but "whether we SHOULD," a question directly relevant to organizational AI adoption.

VSD faces documented challenges in practice, however. Manders-Huits (2011) identifies methodological gaps including unclear stakeholder identification processes, undetermined concepts of "values," and lack of explicit ethical theory for resolving value trade-offs. Practitioner interviews reveal implementation barriers including miscommunication between stakeholders and developers, difficulty translating stakeholder needs "into the machine learning project," and lack of interdisciplinary teams (Sadek et al., 2024).

The Mission-First Framework addresses several of these limitations through specific adaptations (see Table 1). By requiring organizations to articulate a single "mission core" before any technology evaluation, it provides explicit ethical prioritization rather than attempting to balance multiple values simultaneously. By defining "hard boundaries" as non-negotiable constraints, it offers concrete operationalization rather than abstract principles. And by locating authority within existing organizational governance structures rather than requiring new technical expertise, it makes VSD principles accessible to resource-constrained organizations.

2.3 The Mission-First Framework

This paper proposes a three-stage approach to prevent capability-driven value drift:

1. Mission Core. First, articulate the non-negotiable prime directive—the single value that, if lost, makes the organization pointless. This is not a general mission statement but rather the irreducible commitment that defines organizational identity. For a peer support organization, this might be "authentic human connection." For an Indigenous language program, it might be "linguistic sovereignty."

The discipline of identifying a *single* core value forces clarity about what the organization cannot compromise. This reduction involves judgment: organizations often hold multiple values, and selecting one as primary is itself a governance decision. For this reason, the mission core should emerge through participatory deliberation involving organizational leadership, staff, and where appropriate, the communities served. The question to pose is not "What do we value?" but rather "What, if lost, would make us no longer *us*?" The resulting commitment should be documented, ratified through appropriate governance processes, and revisited periodically as organizational context evolves.

Table 1: How Mission-First Relates to Adjacent Frameworks

Framework	Primary Focus	Mission-First Distinction
Value Sensitive Design	Embedding values throughout design process	Prioritizes *single* mission core; formalizes negative space (prohibited uses); designed for adoption decisions, not system design
NIST AI RMF	Risk identification, measurement, management	Adds mission-alignment as prerequisite to risk mapping; emphasizes organizational purpose over technical risk categories
Responsible AI Checklists	Compliance verification across AI principles	Inverts sequence: boundaries before capabilities; focuses on what to *forbid*, not what to *ensure*
Ethics by Design	Embedding ethical constraints in development	Targets organizational governance, not technical development; operates at adoption decision, not implementation

2. Hard Boundaries (The Negative Space). Before listing potential use cases, define the *negative space*: specific applications the AI system could technically perform but is strictly forbidden from performing. These boundaries act as filters for every subsequent technical proposal. Explicitly specifying prohibited uses can be a particularly effective design move in identity-critical settings because it prevents subtle substitution of mission mechanisms with superficially efficient capabilities.

3. The Permissible Sandbox. Only after boundaries are established does the organization explore remaining possibilities. Innovation occurs within the sandbox, and *only* within the sandbox. This constraint can paradoxically enable faster, safer experimentation because the highest-risk applications have already been excluded.

This sequencing maps onto established risk governance frameworks. The NIST AI Risk Management Framework (National Institute of Standards and Technology, 2023) emphasizes that effective AI governance requires organizations to "map" risks before "measuring" or "managing" them. Mission-first adoption operationalizes this principle: the mission core defines what matters; hard boundaries map the unacceptable risks; the permissible sandbox enables measured experimentation.

2.4 Identifying Identity-Critical Contexts

Not all AI deployments require mission-first governance. The framework is designed for identity-critical contexts, operationalized here through five criteria. A context is likely identity-critical when:

1. **Relationship as mechanism of change.** The intervention's effectiveness depends on relational qualities (trust, authenticity, shared experience) rather than information transfer alone. Peer support, pastoral care, and therapeutic relationships exemplify this criterion.

2. **Cultural or epistemic authority at stake.** The domain involves knowledge, language, or practices over which a community claims governance rights. Indigenous language programs, religious education, and cultural institutions often meet this criterion.

3. **Vulnerable populations.** The people affected have reduced capacity to detect or resist harmful AI interactions due to age, mental health status, crisis circumstances, or power differentials.

4. **Identity formation as outcome.** The intervention explicitly aims to shape how individuals or groups understand themselves,

including their values, capabilities, cultural membership, or life direction.

5. **High irreversibility of harm.** Errors or value drift in this domain produce harms that are difficult to undo: damaged trust, cultural loss, reinforced negative self-concepts, or foreclosed identity possibilities.

When multiple criteria are present, the case for mission-first governance strengthens. When none apply, capability-first adoption with conventional risk management may be appropriate. These criteria are intended as heuristics for scoping, not a definitive typology; refinement through application across diverse organizational contexts is anticipated.

3 Methods

This paper presents two **design cases** (Boling, 2010) developed through the author's consulting engagement with each organization. Design cases document the reasoning, constraints, and decisions involved in creating a designed artifact (in this instance, AI governance frameworks).

Data sources include: (1) AI policy documents developed collaboratively with organizational leadership; (2) meeting notes and deliberation records from the policy development process; (3) workflow documentation describing how AI tools were configured and constrained; and (4) reflective practitioner notes on implementation challenges and boundary decisions.

Analytic approach. The cases are presented as illustrative narratives demonstrating the Mission-First Framework in practice. This paper does not claim to evaluate outcomes (e.g., whether mission integrity was preserved over time) but rather to trace the *process* by which organizations translated mission commitments into governance constraints. This approach aligns with reflective practice methodology (Schön, 1983) and design case traditions in HCI.

Limitations of this method. The author participated in developing both frameworks, creating potential bias toward favorable interpretation. The cases were selected for accessibility (existing consulting relationships) rather than theoretical sampling. Claims about framework utility are therefore preliminary; systematic evaluation across independent implementations remains future work.

Organizational consent. Both organizations consented to being named and to the publication of governance framework details. Operational specifics have been abstracted where necessary to protect secure content and avoid disclosing information that could compromise organizational security. No sensitive client records or personally identifiable information are included; the paper reports governance artifacts and workflow constraints at an abstracted level.

Positionality. As an independent researcher, I developed both governance frameworks described here through consulting engagements: with KĀʻEO since 2016 (with the AI governance work beginning in 2025) and with F.E.A.S.T. since 2019 (with AI governance work beginning in 2024). This dual role as framework developer and analyst creates inherent limitations: I cannot claim the detachment of an external observer, and my interpretation of these cases is shaped by my investment in their success. I bring over 20 years of experience in educational technology consulting but no formal training in psychometrics, Indigenous studies, or clinical mental health—domains central to these cases. My perspective is that of a practitioner reflecting on work in progress, not an evaluator assessing outcomes.

4 Design Case A: The KĀʻEO AI Lab

The Kaiapuni Assessment of Educational Outcomes (KĀʻEO) is a Hawaiian-medium assessment used for accountability within Hawaiian language immersion education, operating under strong community expectations around linguistic authority and stewardship (Kūkea Shultz and Englert, 2021, 2023). Because all test content must remain in Hawaiian, every stage of item development, from initial writing to psychometric validation, occurs within a protected linguistic environment.

Table 2: KĀʻEO AI Lab — Boundary Structure

Category	Prohibited Uses	Permitted Uses
Data handling	No KĀʻEO data used for external AI model training; no submission to systems that retain inputs	Document-grounded analysis in ephemeral, access-controlled environments
Content generation	No AI-generated Hawaiian language content published without expert review; no external disclosure of paraphrased content; Hawaiian text remains sole authoritative source	Pattern identification in psychometric data; interpretive hypothesis generation
Human authority	No AI output treated as authoritative without cultural/linguistic expert validation	AI-generated briefs as input to human deliberation

4.1 The Mission Core

The KĀʻEO program provides a psychometrically rigorous, nationally peer-reviewed assessment system for Hawaiian-medium education. This goal requires **linguistic sovereignty**: all test content remains in Hawaiian, and the Hawaiian language community maintains authority over how their language is used and represented. This commitment creates a structural challenge for psychometric analysis. The assessment is written entirely in Hawaiian, but most psychometricians who analyze item performance data do not speak the language. Historically, bridging this divide required either translation of secure content (violating confidentiality and sovereignty) or reliance on a small number of bilingual intermediaries who could span both domains.

This challenge reflects broader tensions in Indigenous data governance. The CARE Principles for Indigenous Data Governance (Carroll et al., 2020) assert that Indigenous communities must have Authority to Control the means of access, use, and interpretation of their data, directly challenging extractive practices common in AI development. The Indigenous Protocol and Artificial Intelligence Position Paper (Lewis et al., 2018) emphasizes that each community "will have its own particular approach" to AI engagement, validating mission-first thinking where organizational values precede technical evaluation.

4.2 The Boundary

A capability-first approach offers an obvious solution: use LLMs to auto-translate secure test items into English for analyst review. Current models are technically capable of Hawaiian-to-English translation. But this application would violate the mission core in multiple ways: it would create an authoritative English "shadow text" of Indigenous knowledge; it would expose secure assessment content to external systems; and it would risk training commercial AI models on Hawaiian language data without community consent.

Applying the Mission-First Framework, the KĀʻEO project team defined explicit hard boundaries in their AI Policy Framework (v4.0). Table 2 summarizes the boundary structure; operational details are omitted for security and privacy.

4.3 The Outcome

With boundaries established, the team identified a permissible sandbox: document-grounded analysis within a controlled workspace. In practice, model use was restricted to a secure environment with organizational access controls and minimized data retention, aligning tooling configuration with mission-defined data handling constraints.

The system was configured to generate "Interpretive Briefs"—summaries explaining *why* items might be underperforming (e.g., "linguistic ambiguity," "cognitive overload," "Depth of Knowledge misalignment")—without translating or reproducing the items themselves. Human reviewers with cultural and linguistic expertise validated

Table 3: F.E.A.S.T. — Boundary Structure

Category	Prohibited Uses	Permitted Uses
Caregiver interaction	No AI for direct support, counseling, or communication with caregivers/families; no caregiver-facing AI of any kind	Human volunteers and staff only for all caregiver contact
Content attribution	No AI-generated content attributed to humans; no undisclosed AI in any external communication	AI drafts clearly labeled; human review and attribution required
Data protection	No sensitive caregiver information entered into AI systems	Anonymized, aggregated data only; no individual case details
Administrative functions	—	Research summarization; internal communications drafting; logistics support

every output before it reached content developers. The governing principle was that humans are not merely checkpoints within an automated process; they *are* the loop (Kūkea Shultz and Brockmann, 2025).

Within the constraints of the policy, practitioners reported that the workflow enabled shorter cycles of interpretive hypothesis generation than prior manual-only processes, without translating or externally disclosing secure Hawaiian-language content. However, the AI's pattern recognition had clear limits: reviewers noted that LLM interpretations of Hawaiian grammatical nuances "lacked the figurative depth and layered meaning characteristic of Hawaiian." This failure validated the necessity of hard boundaries. The AI could identify *where* items underperformed but could not explain *why* within the cultural logic of the language.

By restricting the technology, the program addressed a persistent analytical challenge without compromising its values. The constraint enabled innovation that capability-first adoption would have precluded.

5 Design Case B: F.E.A.S.T.

The KĀʻEO case demonstrates how sovereignty functions as a design constraint. The case of F.E.A.S.T. (Families Empowered and Supporting Treatment of Eating Disorders) demonstrates how **vulnerability** operates similarly.

5.1 The Mission Core

F.E.A.S.T. is a global nonprofit supporting caregivers of people with eating disorders. The organization's mechanism of change is peer support: the lived experience of (typically) parents helping other parents navigate crisis, treatment, and recovery. The mission depends on trust—the knowledge that the person offering guidance has genuinely survived what the caregiver is currently experiencing.

The peer support literature establishes why this mechanism matters. Shalaby and Agyapong (2020) define peer support as "deeply felt empathy, encouragement, and assistance that people with shared experiences can offer one another within a reciprocal relationship," emphasizing mutuality and lived experience as constitutive elements. Information alone is insufficient; the therapeutic mechanism is authentic human empathy shared between people who have walked the same path.

5.2 The Boundary

Current LLMs can generate text that is often difficult to distinguish from human emotional support. They can validate feelings, offer encouragement, provide information about treatment options, and respond to disclosures of suffering with apparent warmth. For a capability-first organization, this might suggest deploying a chatbot to extend support availability and reduce volunteer burnout.

But for F.E.A.S.T., simulated empathy may represent mission failure. For interventions where outcomes depend on the working alliance and

relational factors, substituting a simulated interaction risks undermining a core mechanism of change, even if users report short-term satisfaction with the interaction.

The eating disorder context makes these risks concrete. Sharp et al. (2023), writing in response to the National Eating Disorders Association's chatbot "Tessa" providing harmful dietary advice, state: "Eating disorders are among the deadliest of all mental health conditions" and AI chatbots offering "inappropriate and dangerous information...carries significant risk." If a caregiver in crisis pours their grief into what they believe is a peer conversation, only to later discover they were interacting with a probability engine, the trust that enables the organization to function is fundamentally broken.

The F.E.A.S.T. AI Policy therefore drew an absolute boundary. Table 3 summarizes the structure; operational details are omitted for privacy.

5.3 The Outcome

By explicitly rejecting the LLM's most "human-like" capabilities, F.E.A.S.T. protected its human network. The policy directed AI toward back-office functions: summarizing research literature, drafting internal communications, organizing logistics. These applications may reduce staff burden without touching the therapeutic relationship that constitutes the organization's core value proposition.

The governance framework also established transparency requirements, data protection standards, and cultural alignment tests (all AI applications must support rather than undermine the mission of human connection). The result is a clear operational map: AI serves as administrative infrastructure, never as relational substitute.

6 Broader Implications

The Mission-First Framework addresses immediate organizational governance, but the underlying concern extends further. Two directions warrant attention: the framework's applicability beyond these design cases, and its relevance to emerging questions about AI's role in identity formation.

6.1 Extensions Across Organizational Types

Both cases involve nonprofit organizations with explicit social missions. The framework's applicability to other institutional forms remains to be tested, but several domains present natural candidates:

Religious organizations face analogous tensions when considering AI for pastoral care or spiritual guidance. The mission core ("authentic encounter with the sacred" or "community in faith") may be incompatible with AI-mediated religious counsel, even as administrative applications remain appropriate.

Educational institutions deploying AI for academic advising or career counseling must consider whether the advising relationship is itself educational, whether the process of working through decisions with a human advisor develops capacities that AI-mediated efficiency would bypass.

Healthcare systems adopting AI triage or patient communication tools confront questions about the therapeutic relationship. When does efficiency in routing patients undermine the relational continuity that affects outcomes?

Journalism and civic institutions whose mission cores involve "democratic deliberation" or "public accountability" face distinct boundary questions when AI can generate persuasive content indistinguishable from human-authored reporting.

In each case, the Mission-First Framework offers a common methodology (articulate the core, define boundaries, then innovate within constraints) while the specific boundaries will differ based on organizational purpose.

6.2 The Horizon: Formative AI

The organizational governance challenge examined here may prefigure a broader challenge at the individual level. Early evidence suggests users are already repurposing general-purpose LLMs for guidance, reflection, and identity exploration (Chatterji et al., 2025). Joseph (2025) describes an emerging "Algorithmic Self" where AI systems "do

not passively reflect the self but actively participate in its formation."

This raises questions beyond the scope of the present paper: How do we establish boundaries for technologies whose formative effects emerge through use rather than design? When users exapt productivity tools for life guidance, who is responsible for the resulting identity effects? Can mission-first thinking translate from organizational governance to individual use, and if so, how?

These questions warrant sustained inquiry. The organizational cases examined here may offer preliminary insight: just as F.E.A.S.T. distinguished administrative AI from relational AI, individuals might distinguish instrumental AI use from formative AI use, and establish personal boundaries accordingly. But this remains speculative, a direction for future work rather than a conclusion of the present analysis.

7 Limitations

This framework has several limitations that warrant acknowledgment.

Generalizability. Both design cases involve nonprofit organizations with explicit social missions. The framework's applicability to for-profit organizations, government agencies, or other institutional forms remains to be tested.

Implementation evidence. While both cases document governance frameworks in operation, systematic outcome evaluation (measuring whether mission integrity is preserved over time, whether staff experience the framework as enabling or constraining, whether stakeholders perceive improved trust) requires longitudinal study.

Scope conditions. The identity-critical criteria proposed in Section 2.4 represent initial guidance; refinement through application across diverse contexts is needed. Some contexts may satisfy multiple criteria yet prove amenable to capability-first adoption; others may require mission-first governance despite meeting few criteria.

Power dynamics. The framework requires organizations to articulate a coherent mission core through participatory deliberation. In practice, organizations may have contested values, and the deliberation process may privilege some stakeholder perspectives over others. Integration with established participatory design methods could strengthen this aspect.

Evaluation metrics. The field lacks established metrics for "mission integrity" in AI deployment. Developing measures that can detect value drift before it becomes irreversible, creating early warning systems for the gradual erosion the framework is designed to prevent, represents an important direction for future research.

Failure cases. This paper examines organizations that implemented mission-first governance. Comparative analysis with organizations that adopted AI capability-first, including cases where such adoption produced value drift or mission failure, would strengthen the argument. The NEDA/Tessa incident offers one such case; systematic comparison awaits future work.

Funding. This research received no external funding.

8 Conclusion

This paper proposes a framework and illustrates its application in two design cases. It does not claim to have validated the framework's effectiveness; that requires longitudinal study of mission integrity outcomes across diverse organizational contexts. Rather, it offers Mission-First Adoption as a conceptual tool and a provocation: if AI governance is to serve organizational values rather than erode them, the sequence of decision-making may matter as much as the decisions themselves.

The cases of KĀʻEO and F.E.A.S.T. suggest that defining what AI systems *cannot* do may be as important as specifying what they *can* do. The negative space (the forbidden applications) protects organizational identity from the capability-first trap. For identity-critical contexts (Indigenous language revitalization, mental health support, cultural education, peer-based healing), efficiency cannot be the primary metric. Trust is.

The framework proposed here is deliberately simple: articulate the mission core, define hard

boundaries, then innovate within the resulting sandbox. This simplicity is a feature. Organizations facing pressure to adopt AI rapidly need frameworks that can be implemented without extensive technical expertise. The question is not "What can this technology do?" but rather "What must our organization never become?"

Both design cases suggest that strict governance need not impede innovation; it may enable it. By removing high-risk applications from consideration, organizations can experiment more freely within safe boundaries. The KĀʻEO AI Lab addressed a persistent analytical problem precisely because translation and external training were excluded from the outset. F.E.A.S.T. can explore administrative AI applications without anxiety about mission drift because peer support was declared off-limits before any tool evaluation began.

As LLMs continue to evolve, the boundary between instrumental and formative applications will blur further. Organizations that establish governance frameworks now, before capability-driven adoption becomes entrenched, will be better positioned to integrate these technologies in ways that serve rather than supplant their missions. The alternative is value drift so gradual it becomes invisible until the mission has already been lost.

References

Boling, E. (2010). The need for design cases: Disseminating design knowledge. *International Journal of Designs for Learning*, 1(1):1–8.

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, 16(3):252–260.

Bruder, I. M. (2025). From mission drift to practice drift: Theorizing drift processes in social enterprises and beyond. *Organization Studies*, 46(3):385–407.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., and Hudson, M. (2020). The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):43.

Chatterji, A. et al. (2025). How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research.

Di Troia, S., Parli, V., Pava, J. N., Badi Uz Zaman, H., and Fitzsimmons, K. (2024). Inspiring action: Identifying the social sector AI opportunity gap. Technical report, Stanford Institute for Human-Centered Artificial Intelligence & Project Evident.

Ebrahim, A., Battilana, J., and Mair, J. (2014). The governance of social enterprises: Mission drift and accountability challenges in hybrid organizations. *Research in Organizational Behavior*, 34:81–100.

Friedman, B. and Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

Horvath, A. O. and Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology*, 38(2):139–149.

ISACA (2023). Generative AI 2023: An ISACA pulse poll. Technical report, ISACA.

Joseph, J. (2025). The algorithmic self: how AI is reshaping human identity, introspection, and agency. *Frontiers in Psychology*, 16:1645795.

Kūkea Shultz, P. and Brockmann, F. (2025). Bridging psychometric and content development practices with AI: A community-based workflow for augmenting Hawaiian language assessments. EdArXiv preprint.

Kūkea Shultz, P. and Englert, K. (2021). Cultural validity as foundational to assessment development: An indigenous example. *Frontiers in Education*, 6:701973.

Kūkea Shultz, P. and Englert, K. (2023). The promise of assessments that advance social justice: An indigenous example. *Applied Measurement in Education*, 36(3):255–268.

Lewis, J. E., Arista, N., Pechawis, A., and Kite, S. (2018). Making kin with the machines. *Journal of Design and Science*. Issue 3 (Resisting Reduction series).

Manders-Huits, N. (2011). What values in design? The challenge of incorporating moral values into design. *Science and Engineering Ethics*, 17(2):271–287.

National Institute of Standards and Technology (2023). Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology.

National Institute of Standards and Technology (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, National Institute of Standards and Technology.

Sadek, M., Calvo, R. A., and Mougenot, C. (2024). Challenges in value-sensitive AI design: Insights from AI practitioner interviews. *International Journal of Human-Computer Interaction*, 41(1):10877–10894.

Schön, D. A. (1983). *The Reflective Practitioner: How Professionals Think in Action*. Basic Books.

Shalaby, R. A. H. and Agyapong, V. I. O. (2020). Peer support in mental health: Literature review. *JMIR Mental Health*, 7(6):e15572.

Sharp, G., Torous, J., and West, M. L. (2023). Ethical challenges in AI approaches to eating disorders. *Journal of Medical Internet Research*, 25:e50696.

Singla, A., Sukharevsky, A., Yee, L., Chui, M., and Hall, B. (2024). The state of AI in early 2024: Gen AI adoption spikes and starts to generate value. Mckinsey global survey on ai, McKinsey & Company.