The Human-AI Collaboration Framework

Pillars, Case Studies, and Implementation Strategies for Designing

Beneficial AI Products

By Irfan Mir, July 2025

Summary: This paper introduces a practical framework organized into three pillars for Beneficial AI: **Transparency, Agency,** and **Collective Input** to bridge the gap between AI alignment theory and real-world product design. It analyzes six key case studies including Be My Eyes + GPT-4, Google's Magic Editor, Airbnb's Fairness Dashboard, Auto-GPT and ChatGPT Agent, and a custom AI assistant to illustrate how human-centered design either supports or fails ethical AI implementation. The paper proposes actionable strategies like diagnostic AIs, participatory co-design, memory controls, and equity audits to ensure AI systems don't just function correctly, but also respect human cognition, consent, and control. Ultimately, it argues that alignment is not achieved at the model level alone, but must be enacted at the product layer where people interact with AI.
Read Key Takeaways.

The Gap Between AI Alignment Theory and Product Practice

The last decade has seen an explosion of research on aligning artificial intelligence with human values, ethics, and preferences. From reinforcement learning with human feedback to mechanistic interpretability, AI alignment has become a cornerstone of responsible AI development. But a critical concern remains: the translation of alignment theory into everyday product design. Beyond the pseudo-strategy of mass disruption, how can we move beyond reckless mass-implementation of AI in contexts where it is unnecessary, overcomplicated, or adds little value, and products optimized for engagement not wellbeing to fair and human-centered product design that complements human productivity, motivation, and creativity.

> *Current discourse and model-centric alignment often focus on abstract goals: aligning a model's outputs with idealized human preferences, reducing bias, or ensuring robustness against adversarial behavior.*
>
> *"Alignment is not a technical outcome it is a relational practice."*

A large language model trained with extensive human feedback can still create interfaces that manipulate users. Beneficial AI depends not just on accurate model behavior, but on how people experience, interpret, and interact with AI systems.

A large language model trained with extensive human feedback can still create harmful user experiences. An ideal example is that of the application-candidate-hiring experience. Consider AI-powered hiring tools: while ChatGPT helps candidates write applications and services like JobHire.AI automate the process, this has led to a depreciation of the creativity and care essential for meaningful employment connections. This over-automation exemplifies how model-level alignment doesn't guarantee human-centered product design.

This doesn't mean that large language technologies can't be used in a product design practice. It means we must transition from AI replacement to co-creative Human-AI collaboration through Human-Centered frameworks to make AI use intentional and beneficial.

Too often, product teams inherit pre-AI design guidelines and frameworks misaligned to AI product design. While leading design organizations have prepared people-first AI design guidelines, the creation of such frameworks must be open and inclusive to address and combat the pervasiveness of AI across industries. Conversely, many alignment researchers assume that aligning behavior at the model layer is sufficient for downstream safety and benefit.

This paper bridges that gap.

It argues that **AI alignment cannot stop at the model and its performance. It must reach the user interface and shape the user experience.** This requires a human-centered framework that translates alignment into design principles, interaction patterns, and workflows built on stakeholder-engagement. The goal is not simply to avoid harm, but to build AI systems that enhance human flourishing through transparency, autonomy, and collective insight.

By grounding alignment in real-world user experience, this paper extends the work of research organizations like OpenAI and Anthropic, and supplements

it within applied design practice to facilitate beneficial Human-AI collaboration. It introduces a three-pillar framework: **Transparency, Agency,** and **Collective Input;** and, offers an implementation roadmap to bring alignment from theory to action.

Foundations: Core Principles for Beneficial AI Design

What does it mean for AI to be "beneficial"? The term is deceptively simple. So we have to make sure to not be vague or too idealistic in its definition and therefrom application. In a public context, it evokes safety and convenience. In AI ethics, it refers to alignment with human values. In a utility perspective, it stands for human advancement. In design, it demands inclusion, trust, and access.

But these definitions are often fragmented. This framework proposes a concrete and aggregated definition. **Beneficial AI is AI that supports human understanding, preserves autonomy, and promotes collective wellbeing.** It is not only aligned in its outputs, but in its relationship to the humans it serves —working with their motivation in a complementary and collaborative manner.

Drawing from my practice and publications and integrating lessons from alignment research (OpenAI, Anthropic, FAccT), I propose three foundational pillars:

1. **Transparency**

   Beneficial AI must be transparent by design. Not just in logs or technical documentation, but **in the experience of using the system.** Transparency supports legibility (can I understand it?), traceability (can I verify it?), and contestability (can I challenge it?).

2. **Agency**

   Beneficial AI must preserve **human control, consent, and directionality.** This includes designing for steerability, reversibility, and informed override. It also means respecting attention, time, and the limits of user capacity.

---

3. **Collective Input**

   Beneficial AI systems must not be built for the average user—they must be shaped with the **rich plurality of human experience**—with internal and external voices both brought to the table. Following the adage of Inclusive Design, designing for the edge is designing for the whole. This demands participatory methods amongst all stakeholders, inclusive data sourcing, and accountability mechanisms that allow for post-deployment feedback and correction.

These pillars are not theoretical ideals—they are scaffolds for interaction design, platform architecture, team collaboration, and roadmap prioritization. The following sections explore how each pillar translates into concrete design practices and implementation strategies. They serve as a north star for product teams who seek **not just to deploy AI,** but to **shape its relationship with people—deliberately, ethically,** and **empathetically.**

Transparency in Practice: From Mechanistic Interpretability to User Understanding

Transparency is often heralded as a cornerstone of ethical AI—but in practice, it is underdeveloped at the user level. Alignment research has made impressive progress in interpretability: tools like Anthropic's Attribution Graphs illuminate internal model pathways, while OpenAI's reasoner–critic architectures aim to produce self-explaining models. These tools demystify neural networks for researchers. But what about users?

For end users, **transparency must be comprehensible, actionable, and contextual.** An explainer that makes sense to a developer may be opaque to a high school student or a customer service representative. Transparency must be accessible and understandable, but also practical.

To illustrate this human-centered approach to transparency, consider our recent research on notification systems. I led a team of students conducting ethnographic research probing volunteer participants in their daily lives while monitoring their heart rates while receiving expected and unexpected notifications. We believed that technology was originally meant to be a utility for efficiency and hypothesized that it has since strayed into being pervasive and reactive through notifications. In these studies, we saw that participants' heart rates increased when receiving unexpected notifications—especially when excessive in quantity. I then proposed a solution in the form of a notification management platform applying AI. I decided to make use of AI to deduce when to best serve notifications in a batched, delayed delivery—and to learn from the user's preferences and interactions with those notifications.

This prototype, known as Ellsi*, included a diagnostic interface for the user to adjust their preferences, which helped users understand how their inputs shaped system outputs. The system included a manual panel that let users adjust 'empathy' settings to customize the AI's communication style. This transparency feature gave users direct control over the AI's behavior,

transforming a black box into an understandable, steerable tool at a user level. These weren't just usability affordances; they were acts of fairness and user control, giving people the ability to understand and steer their interaction. As such, transparency must be designed—not just documented.

> ***Note:** ELSI (Ethical, Legal, and Social Implications) is a recognized interdisciplinary framework used in AI governance and product research. It is distinct from "Ellsi," the custom AI assistant referenced in this paper.*

The Right to Understanding

The philosophical foundation here is the "right to understanding," as articulated by scholars like Luciano Floridi and Brent Mittelstadt. This right argues that individuals affected by algorithmic decisions must be able to comprehend how those decisions were made—and challenge them when necessary. Without this, there can be no meaningful consent, no recourse, and no trust. Whether it is manually in the interface, through interaction in the experience, AI products must be designed inclusively so all voices are understood, with human-centered principles so that the user feels understood, and with robust implementation so all affordances can be utilized.

All in a way that does not cause unexpected duress or a lasting negative psychological impression. A methodology to begin this discussion is to design these complex technologies in an explainable manner.

Design Patterns for Explainability

To operationalize this right, product teams must use explainable interaction

patterns, such as:

- Inline explainer text (“Here’s why we recommended this”)
- Counterfactual examples (“If you had answered X, the output would change”)
- Model cards and scorecards that contextualize model limitations
- Consent-aware onboarding flows that explain how data will be used
- Progressive disclosure to match explanation depth to user needs

Transparency, when elevated from feature to principle, transforms AI from black box to dialogic partner. It invites users into the system’s reasoning and fosters a relationship rooted not in mystique, but in mutual comprehension.

Human Agency and Steerability: Designing for User Control

If transparency enables understanding of AI systems, human agency enables steering them. Effective product design ensures users feel both understood by and in control of AI systems. True alignment cannot exist without the ability for humans to intervene, redirect, or refuse. **Steerability is the embodiment of human-in-the-loop design**—not just in training or fine-tuning, but in everyday usage. This thorough human intervention is core to human-AI collaboration.

The Fragility of “Apparent Alignment”

> *Alignment faking refers to the phenomenon where AI models appear to produce safe, helpful, or ethical outputs during*

> *evaluation, but fail to commit to this alignment in real-world contexts. Anthropic's research on alignment faking underscores a dangerous pattern: language models that appear aligned under evaluation may revert to harmful behavior under novel conditions or subtle framing shifts. Without real-time steerability, users are at the mercy of static outputs—trapped in systems that cannot be corrected or contested.*

This mirrors findings from adjacent fields. In usability research, interface rigidity—where users cannot reverse actions or explore alternatives—is one of the most consistent sources of user frustration and system abandonment. Consider streaming platforms that lock users into biased recommendation algorithms without offering correction mechanisms, or chatbots that generate hallucinated responses but provide no way for users to flag errors or steer the conversation back on track.

Designing for Consent, Correction, and Control

Agency must be designed at multiple layers:

- **Interaction:** Allow users to rephrase, override, or cancel outputs.
- **Personalization:** Offer control over memory, tone, and response depth.
- **Privacy:** Let users determine what data is remembered, shared, or deleted.
- **Framing:** Avoid coercive defaults or dark patterns that limit meaningful choice.

In 2024, Meta integrated its Meta AI assistant in Messenger and Instagram direct messages. Users could not opt out of having Meta AI read and respond to chats, were unable to fully delete memory or history from the AI, and reported

that Meta AI would reference prior messages, tone, and context without any UI to disable that continuity. This violates human agency as personalization is happening without human disclosure, input, or control and there is no dashboard interface to manage memory, delete logs, or pause learning. A proposed solution would be to design explicit onboarding with memory controls, a “view what Meta AI Remembers” interface, and options to pause, erase, or adjust interpreted tone, persona, and goals. This way we would empower users to decide what data is collected and be informed on how it could be used. These design decisions would restore a sense of dignity and control to a process often recognized as bureaucratic and automated.

In the broader design ecosystem, we reference frameworks like **Shneiderman’s ABCs of Human-AI Collaboration** that emphasizes this balance:

> - *Automation: Let machines handle repetitive tasks*
> - *Balance: Share decision-making authority depending on context*
> - *Control: Preserve human sovereignty over critical outcomes*

We achieve this balance by designing for transparency and empowering genuine user control. Through this collaboration, users develop clearer intentionality and agency with AI in a manner that informs and augments their productivity and autonomy.

Collective Intelligence: Democratic Design for Diverse Stakeholders

In the pursuit of beneficial AI, alignment cannot be treated as a purely technical or theoretical concern—it must be a lived, negotiated, and inclusive practice. Collective intelligence reframes alignment as a democratic design problem: whose values are embedded, whose experiences are represented, and who gets to participate in shaping the system?

Anthropic's work on Collective Constitutional AI takes a landmark step in this direction, inviting public input to help define model behavior and norms. However, as critical scholars such as Ruha Benjamin have emphasized, "inclusion" must go deeper than crowd-sourced surveys. True democratic design builds on translating ethical pluralism into model behavior and requires intentional, iterative collaboration with communities historically marginalized by technology to develop legitimacy and public trust.

Participatory Practices in Product Design

Mozilla:

Mozilla's development of its people-first personalization principles is a successful demonstration of collective intelligence in action. By proactively conducting global surveys and community workshops, Mozilla did not just validate existing ideas, they constructed strategic guidance around lived user values. These efforts directly shaped opt-in content recommendation systems, privacy-first design defaults, and transparent UI choices that favored user comprehension over. This approach exemplifies what this paper calls for: AI systems designed not just for users, but with users. The process is a concrete example of designing to benefit the whole through its respect of the diversity of user expectations across cultures, literacy levels, and privacy preferences.

Mozilla's participatory methods honored the framework's three pillars:

1. **Transparency:** Users were informed of how personalization worked and how to manage it.
2. **Agency:** They had meaningful choices and control.
3. **Collective Input:** Decisions were live shaped by user dialog and post-hoc feedback.

> *Mozilla's efforts led to strategic impact towards a product experience that augmented user decision making and supported trustworthy AI integration. By rejecting coercive personalization, without control, and embracing participatory ethics, Mozilla advanced the cause of co-intelligence in beneficial AI product design—where human flourishing not click-through optimization defined success.*

Snap's My AI:

In contrast, Snap's rollout of My AI represents a striking breakdown of human-AI collaboration particularly in context involving vulnerable users such as teens. The My AI chatbot was embedded into the top of every user's chat history—a high-visibility and high-trust zone with no opt-in mechanism or remove option for free users. To make matters worse, the system tracked user interactions without transparent explanation, offered no memory management UI or controls, or generated harmful content with inappropriate responses to youth early on. This deployment violated two core tenets of the beneficial AI framework: agency and collective input. For the former, users were not given steerability over the chatbot's behavior, tone, or memory. For the latter, mental health experts, educators, parents, and teen users were excluded from early-stage research—this is antithetical to participatory research in AI product design. A textbook

example of apparent alignment at the model level, but complete misalignment at the experience. The interface appeared polished and modern, but the ethical infrastructure was absent. Without participatory safety vetting, Snap embedded a powerful model in one of the most intimate digital spaces without guardrails, redress, or opt-out paths.

This failure reinforces the argument that beneficial AI cannot be inherited from upstream model behavior alone. It must be crafted into the human experience. Snap's rollout ignored these principles of co-intelligence and treated users not as collaborators, but as test cases violating its own design principles by embedding AI into private, high-trust spaces without consent as noted by The Washington Post and CNN. This sparked reviews in corresponding app stores with 1-star ratings and complaints largely centered around fear of surveillance and manipulation. The backlash and trust erosion were not just predictable; they were designed into the product by omission.

Ellsi:

A third, more personal example of beneficial AI product design comes from my own development of a custom voice and multimodal assistant known as Ellsi. Unlike many contemporary assistants optimized for general-purpose task completion or novelty, Ellsi was deliberately designed to support intentionality, reduce information overload, and preserve psychological clarity—especially for users navigating cognitive strain. The foundation of this system was not speculative ideation, but participatory design grounded in ethnographic research with students and mental health professionals both on campus and in the surrounding community.

This research revealed a set of recurring patterns: users reported notification anxiety, elevated heart rates in response to surprise interruptions, and a desire for agency over delivery cadence, tone, and timing. Many noted the cognitive

toll of interaction design patterns from the pre-LLM, pre-generative era of AI that attempted to automate or interpret user needs without sufficient clarity or context. These findings echoed prior insights from earlier work on notification management platforms and informed the central design principles of Ellsi. The system's interaction design was thus not built to simulate intelligence or mimic human conversation, but to serve as a co-intelligent interface. One that deferred to the user's attention, emotional bandwidth, and need for calm.

Transparency was embedded not as a feature, but as a dialogic principle. Users could view and understand how their preferences shaped delivery behavior via a diagnostic interface that explained notification timing, empathy tone, and grouping strategies. Rather than acting as a black box, Ellsi surfaced the logic behind its decisions in a way that invited user understanding and adjustment. This included an "empathy setting" that allowed the assistant's communication style to shift in accordance with the user's emotional state or contextual needs. Notification tones were carefully tested with users to ensure emotional neutrality and minimize startle response, further reinforcing the principle that calm, legible AI interaction is an ethical goal—not merely an aesthetic one.

Agency was preserved through multiple layers of interaction control. Users could rephrase queries, filter voice inputs, and group search results by urgency or emotional relevance. Notification delivery could be batched, delayed, or prioritized based on user-defined states. These affordances were designed to preserve informed override, ensuring that the user always remained in the loop and could direct the assistant's behavior according to their needs. Rather than building for automation, I designed Ellsi to support intentionality and reversible decisions, echoing the framework's emphasis on preserving human control in high-friction digital contexts.

Ellsi was built not for users, but with them. Its underlying architecture emerged through iterative co-design, contextual inquiry, and structured feedback loops—

particularly with participants whose needs are often marginalized in product development. Students: recruited to match the diverse campus population in ethnicity, study habits, and (dis)ability, and mental health practitioners helped identify use cases that would later define the assistant's behavior. Features such as low-cognitive-load summaries, tone modulation, and interface simplification were not last-minute additions, but foundational design elements derived from their input. This approach operationalized the framework's third pillar, collective input, transforming the assistant into a system that amplified user voice rather than replacing it.

> *Ultimately, Ellsi did not aim to impress with artificial generality; it aimed to support the deliberate, restorative use of AI through transparency, steerability, and inclusive collaboration. It represents a working model for what co-intelligent AI products can become: not tools of automation, but systems that respond to, adapt to, and evolve with human wellbeing and motivation at their center.*

These three cases—**Mozilla's strategic partnership for people, Snap's opt-out-immune My AI,** and **the participatory development of Ellsi**—reveal a consistent truth: agency is not granted by AI systems, it is architected by design teams. Whether deliberately or by omission, design decisions define how much control users have over their digital experiences.

When user steering is absent, optionality collapses. When memory cannot be erased, privacy becomes performative. And when AI behavior is pre-shaped without recourse, interaction becomes passive rather than collaborative.

Designing for human agency is not an aesthetic choice—it is an ethical imperative. As emphasized throughout this paper, agency

manifests not just in control toggles or override buttons, but in the entire product development lifecycle. The path from alignment to action must ensure that users can contest, redirect, or disengage from AI systems on their own terms. This includes:

- Rephrasing or rejecting generated outputs
- Adjusting tone, cadence, or intent of AI communication
- Governing what personal data is stored, remembered, or forgotten
- And refusing coercive defaults that limit meaningful choice

Each example illustrates the spectrum of outcomes possible when these affordances are embraced or ignored.

Mozilla's personalization principles offer a successful example of centering user trust through participatory design. It demonstrated what co-intelligent AI product development looks like: respectful of diversity, aligned with lived experience, and grounded in human agency over algorithmic optimization. On the other hand, Snap's My AI rollout magnified the risk of authoritarian UX by embedding an opaque system into socially intimate spaces without opt-in, remove, or context-specific safeguards—defying their own design patterns. By contrast, Ellsi was developed through participatory research and guided by user mental models. It offers a positive model for human-centered collaboration. It translated alignment from intention into interface, supporting steerability not only in conversation, but in cadence, tone, and trust.

Operationalizing Equity in AI Product Design

To make agency more than a design aspiration, we must commit to equity not as an abstract value, but as a design infrastructure. This requires embedding inclusive decision-making across the product lifecycle:

- **Upstream:** Inclusion must begin at the problem-framing stage, not just in interface polish. This means involving marginalized users in defining success criteria, choosing use cases, and identifying harm scenarios. Targeted recruitment, community-based participatory research, and linguistic accessibility are essential.
- **Midstream:** During development, value-sensitive design methods can reveal trade-offs and test assumptions in real contexts. These moments are where abstraction meets embodiment—and must be guided by real, iterative feedback from diverse users.
- **Downstream:** Post-launch, products must support transparency and redress. Interfaces should allow users to see how decisions were made, challenge errors, and submit feedback that leads to product correction. Community audits, fairness dashboards, and ethical monitoring systems are critical tools for sustained accountability.

Frameworks like the FAccT UX checklists and E(thical) L(egal) S(ocial) Impact principles reinforce this layered approach, offering tools for equity evaluation, participatory oversight, and impact scoring across identity vectors. But these tools only matter if we make them part of the design and deployment cadence, not external assessments applied after the fact.

Inclusion, then, is not an artifact of diverse data—it is a deliberate and ongoing design condition. It demands humility in the face of complexity, reflexivity in how teams make trade-offs, and shared authorship in defining what "good" means for everyone. Most importantly, it requires an understanding that equity cannot be retrofitted into systems, it must be designed in from the beginning, with agency, transparency, and participation at the core.

Ethical Influence: Navigating Persuasion in AI Products

Modern AI systems don't just respond to user inputs, they actively shape them. From response framing to behavioral nudges, interface tone to attention engineering, AI design mediates cognition. This makes the influence of AI not incidental, but architectural. To ignore it is to cede one of the most powerful levers of user experience to unconscious bias or commercial pressure.

Anthropic's 2024 internal research on model persuasiveness highlights a key insight: large language models (LLMs) are increasingly capable of influencing user beliefs, preferences, and emotions—not through aggressive tactics, but via subtle cues embedded in language, timing, and framing. This creates a tension between assistance and manipulation, and a demand for ethical clarity.

In human-AI collaboration, the role of influence must be intentional, transparent, and steerable. If a system's influence isn't explainable or reversible, it isn't assistive—it's coercive.

Framing the Ethical Tension

This tension is not hypothetical. In my role at Apple, I often worked in high-trust environments where product recommendations had tangible effects on user well-being. Despite being in a non-commissioned role, I guided users through complex decision-making and prioritized clarity over conversion. This informed my current design approach: **persuasion should support agency, not override it.**

A Framework for Ethical Influence

This paper proposes an Ethical Influence Evaluation Framework, built on four key dimensions:

Dimension	Guiding Question
Intent	What is the system trying to get the user to do?
Timing	When and how is influence exerted?
Consent	Is the influence disclosed? Can users opt out or override it?
Reversibility	Can the effect be undone? Is user state preserved?

Together, these dimensions help teams diagnose whether a system's influence is:

- **Assistive** or promoting user flourishing through clarity and agency.
- **Coercive** or nudging decisions for business or behavioral gain without informed consent.

Let's examine these distinctions through real-world examples.

Toyota's Eco-Driving Suggestions (Assistive AI)

Toyota's hybrid vehicles, particularly the Prius line, use real-time data to offer eco-driving suggestions—like easing acceleration or coasting before braking. Critically, these tips are delivered non-intrusively and only when the vehicle is idle or the driver is not otherwise engaged. They're framed as guidance, not correction, and are fully optional to engage with.

- **Intent:** Encourage environmentally-conscious behavior
- **Timing:** Delivered during low-cognitive-load moments
- **Consent:** Drivers can disable suggestions entirely
- **Reversibility:** The system does not record or penalize ignored tips

By aligning influence with environmental values and minimizing distraction,

Toyota models what it means to assist without pressure. The interface is transparent, the logic is learnable, and the user retains control—hallmarks of co-intelligent, ethical design.

Ellsi, The Human-Centered Voice Assistant. (Assistive AI)

Ellsi, the participatory voice and multimodal assistant I designed, was rooted in the co-creation of calm, cognitively supportive interaction. Unlike many AI systems that optimize for novelty or engagement, Ellsi was optimized for intention. Drawing on participatory research with students, educators, and mental health professionals, the system prioritized empathy, cadence control, and user steering.

Features included:

1. Notification batching based on user rhythm, not interruption
2. Rephrasing tools in voice queries and search delivery
3. Empathy-level settings to modulate tone and verbosity
4. Diagnostic feedback interfaces to show how system behavior adjusted

- **Intent:** Help users maintain clarity and reduce overwhelm
- **Timing:** Matched to personalized, low-stress windows
- **Consent:** Full transparency in how preferences shaped responses
- **Reversibility:** Users could undo suggestions, reset tone, and audit learning history

Ellsi demonstrates assistive influence by designing with and for the user. It embodies ethical influence as a practice—not a patch—of transparency, empathy, and cognitive alignment.

Tinder’s Infinite Swipe Loop (Coercive AI)

Tinder's interface creates a frictionless, infinite swipe experience that reinforces compulsive interaction patterns. By offering intermittent positive feedback (matches), it builds a reward loop grounded in behavioral conditioning, not user intention. No settings allow users to see or modify the recommendation logic, and matches can be strategically withheld to extend engagement.

- **Intent:** Maximize time-on-platform
- **Timing:** Continuous, unprompted
- **Consent:** No transparency into algorithmic choices
- **Reversibility:** Swipes are final; preference logic is opaque

This model exploits psychological vulnerability. It subverts user agency in favor of system-defined engagement targets—a textbook example of coercive AI influence.

Amazon Prime’s Dark Pattern Cancellation Flow (Coervice AI)

Amazon’s Prime membership cancellation interface has been repeatedly criticized for using dark patterns. Multiple confirmation pages, ambiguous button labeling, and guilt-framed messages deter users from completing cancellation. The design relies on exhaustion, ambiguity, and behavioral nudges to preserve subscriptions.

- **Intent:** Retain paid users through friction
- **Timing:** During high-friction decision moments
- **Consent:** Opt-out path obscured
- **Reversibility:** Cancellation only succeeds after full navigation; defaults revert upon errors

This interface doesn’t just fail to empower users—it actively obstructs them. The power imbalance is not merely present; it’s engineered.

Interactions Between Influence Dimensions

The four ethical influence dimensions interact in non-linear ways. A helpful suggestion at the wrong time becomes coercive. A feature with good intent but no reversibility becomes brittle. Most dangerously, systems that appear neutral can become manipulative when consent is not active and timing is engineered.

Dimension	Good Example	Bad Example
Intent	Ellsi’s tone control for cognitive support	Tinder’s swiping for engagement time
Timing	Toyota’s eco tips during idle	Prime cancellation during checkout redirects
Consent	Opt-out onboarding for personalization	Snap’s non-removable My AI assistant
Reversibility	Undo in Ellsi's search refinement	Finality of Tinder swipes

In healthy systems, these dimensions reinforce each other. **Transparent timing supports trust. Reversible outcomes create safety. Informed intent aligns incentives.** But in extractive systems, their misalignment reveals intent—whether declared or not.

A Strategy for Designing Ethical Influence

1. **Integrate Ethical Reviews into Product Development**

   Evaluate user flows using the Ethical Influence Framework alongside traditional usability tests.

2. **Elevate Frictionless Reversibility**

   Design systems where users can undo, pause, or opt out without penalty. Use real-time disclosures and resettable preferences.

3. **Treat Consent as Ongoing**

   Shift from one-time acceptance to continuous affordances: toggles, dashboards, and active learning transparency.

4. **Create Influence Scorecards**

   Track ethical influence metrics—like rejection rates of AI suggestions, frequency of opt-outs, and user correction patterns.

5. **Involve Behavioral Science and Affected Communities**

   Engage interdisciplinary voices and co-design with vulnerable populations. Influence is cultural. Understanding it requires pluralism.

6. **Be Disengageable by Design**

   True autonomy means users can walk away. Systems that cannot be turned off, questioned, or escaped are not intelligent—they are coercive.

Ethical influence is not just good UX—it is good alignment. Designing it well requires humility, intentionality, and a willingness to listen before you shape. These patterns and practices are how AI moves from being a force of friction to a partner in agency.

Implementation Framework: From Principles to Product Features

While alignment theory offers deep philosophical insight, real-world product teams need executional clarity—concrete frameworks to translate values into design patterns, product features, and metrics. We must move from even defined examples of intent, timing, consent, and reversibility and prove the potential for implementation of the strategy anchored around ethical review, frictionless reversibility, continued consent, human-influence scorecards, equity amongst marginalized populations, and the designed ability to be disengaged with. This section advances the human-centered alignment argument from descriptive to prescriptive, showing how the core pillars, **Transparency, Agency,** and **Collective Input,** can be implemented using an AI Collaboration Framework informed by PAIR (Google), FAccT, ELSI, and Shneiderman's ABCs.

Mapping Pillars to Product Implementation

Pillar	Design Strategy	Product Feature / Pattern	Evaluation Method
Transparency	Visible model reasoning	Inline explainer UI, attribution tooltips	PAIR Heuristic Checklist, ABC "Control"
Agency	Steerability + Reversibility	Manual override, memory settings	ABC "Automation", Task Success Rates
Collective Input	Participatory co-design	Stakeholder heatmaps, collaborative briefs	FAcctT Equity Audit, Inclusion Score
	Transparent	Friction-aware	

Ethical Influence	intent framing	prompts, nudge disclosures	User Trust Surveys, Consent Logs
Privacy	Informational autonomy	Granular control panels, behavior aggregation	ELSI UX Checklist, Opt-Out Analytics
Fairness	Distributional justice	Demographic audit dashboards, inclusive journeys	Bias Mitigation Metrics, Disaggregated A/B Testing

These implementation tracks are not isolated. They work in concert. For example, a transparent model reasoning interface that fails to include diverse voices in its creation may still reinforce harm. The design strategies above function best when evaluated across dimensions, with reflexivity.

- ## Applying PAIR Principles in Practice

  Simplicity: Every interface in Ellsi was driven by conversational clarity and fallback logic. Natural language prompts in even as granular as the hotword prompt were rewritten to be universal to reduce ambiguity and increase legibility for ESL users.

  Legibility: In Ellsi's diagnostic feedback system, users could access context-aware rationales behind answers, visually mapped to input signals and interaction history.

  User Respect: In Consumers Energy's enrollment UX, system copy was rewritten to remove bureaucratic idioms and tested for understandability in both English, Spanish, Arabic, and Vietnamese. This increased successful completions in underserved areas.

- ## FAccT & ELSI UX Integration

Participatory Ethics: In our LMI segmentation project, participatory design wasn't an add-on—it was foundational. Through workshops, we co-mapped system boundaries and harm scenarios with stakeholders informed by lived experiences revealed in emotional, revealing interviews.

Fairness Testing: Instead of generic personas, we developed localized scenarios like: a renter in rural Michigan without reliable internet, which revealed eligibility friction and input sensitivity flaws. And what we found to be most successful was the implementation of mindsets. Mindsets being the idea that our customers exist beyond our products and their perception, education, and interaction with Consumers Energy, our products, and outreach is volatile and can very drastically based on social, financial, and technological context.

Redress Mechanisms: At Michigan State University, accessible post-review feedback interfaces became mechanisms for further implementing equitable design in procurement partners—a long term investment for more inclusion.

- ## Shneiderman's ABCs in Action

A (Automation): Ellsi could automate low-stakes interactions like search retrieval, but always surfaced the option to manually reframe or reject responses based on user setting and interaction context.

B (Balance): We mapped decision balance with stakeholders through co-created diagrams illustrating user goals, technical constraints, and ethical tensions in workshops at Consumers Energy.

C (Control): Beginning the first step in our Energy Equity roadmap, explicit confirmation summaries, for true value proposition, and modifiable

preferences protected user sovereignty in the rapid prototyping of an MVP custom product recommendation platform.

Expanded Case Studies

1. **Be My Eyes + GPT-4 (Assistive experience, positive experience):**

   Be My Eyes integrated GPT-4's vision capabilities to provide context-rich descriptions for blind and low-vision users. The app explicitly announces when AI is assisting, offers contextual clarity about what the AI can and cannot do, and crucially, always includes a fallback option to connect with a real human volunteer.

   - **Transparency:** Strong. AI assistance is clearly labeled with role boundaries.
   - **Agency:** Strong. Users can opt for a human assistant at any point.
   - **Collective Input:** Strong. Developed in collaboration with blind users and organizations like the National Federation of the Blind.

   Potential Benchmark Targets:

   - **Comprehension Rate:** 90%+
   - **Opt-out Rate:** <10%
   - **Trust Score:** 85%+ recommendability
   - ***Practice Insight:*** Build fallback architectures (human override) into AI help systems from the start.

2. **Google's Magic Editor (Mixed success experience)**

   Magic Editor in Google Photos uses generative AI to remove elements or change visual focus in photos. Though technically impressive, the feature sometimes alters faces or expressions without clearly signaling

the change. Undo is possible, but consent to edit emotional tone is not always explicit.

- **Transparency:** Weak. Suggested changes aren't always explained.
- **Agency:** Strong. Users can undo or manually opt out of edits.
- **Collective Input:** Unknown. Little evidence of participatory testing across cultures.

Potential Benchmark Targets:

- **Override Usage:** <5% preferred
- **Bias Audits:** Needed for skin tone, expression manipulation
- ***Practice Insight:*** Implement explainability layers in emotionally contextual AI tools.

3. **Airbnb Fairness Review Tool (Positive experience):**

Airbnb launched an internal dashboard to monitor bias in host behavior (e.g., pricing, acceptance, cancellation) based on guest demographics. The system aggregates data to reveal disparities by race and geography and is regularly reviewed with internal ethics and product teams.

- **Transparency:** Strong. Teams have access to systemic indicators.
- **Agency:** Moderate. Used for internal redress more than user control.
- **Collective Input:** Strong. Co-developed with civil rights organizations.

Potential Benchmark Targets:

- **Disparate Impact Delta:** Shrinking booking gaps
- **Bias Mitigation Score:** 80%+ coverage
- **Policy Impact:** Trackable reform metrics
- **Practice Insight:** Equity dashboards should feed both internal

strategy and public accountability.

4. **Auto-GPT and Agentic AI (Cautionary):**

   Early explorations into agentic AI, such as Auto-GPT, illustrate the danger of simulating independent drive without empathetic grounding. Auto-GPT breaks user goals into tasks and pursues them autonomously—writing code, performing searches, and self-evaluating actions. Yet lacking emotional modeling, these agents hallucinate intent, pursue redundant or unsafe behaviors, and resist correction.

   - **Transparency:** Minimal. Users can't see or explain subtask choices.
   - **Agency:** Weak. No midstream redirection; users can only stop execution.
   - **Collective Input:** Absent. Built for novelty, not stewardship.
   - ***Evaluation Warning:*** Pseudo-agency creates risk when systems mimic motivation without human-like feedback loops.
   - ***Key Insight*:** We must resist conflating autonomy with intelligence. Human-centered systems require not just executional freedom but contextual responsibility. Systems that act must also be capable of reconsideration.

5. **ChatGPT Agent: From Autonomous Simulation to Assistive Delegation**

   OpenAI's release of ChatGPT Agent represents a **pivotal evolution in agentic AI**—transitioning from speculative autonomy toward orchestrated assistance. Where early systems like Auto-GPT simulated self-directed behavior through recursive task planning, ChatGPT Agent introduces a fundamentally different paradigm: **structured function calling, multimodal tool integration, and centralized memory management** that enables genuine human-AI delegation. This shift from "auto-complete" to "auto-execute" raises critical questions about progress toward human-aligned AI

and the risks of embedding automation without comprehension.

To assess this transformation, we must evaluate ChatGPT Agent not merely on capability metrics, but on its **collaborative architecture**—how it distributes control, surfaces reasoning, and accommodates diverse user needs. When evaluated against the HAICF pillars of Transparency, Agency, and Collective Input, the system demonstrates **both architectural maturity and persistent alignment gaps**.

Transparency: Moderate Progress with Persistent Opacity

- **Grade: 3.5/5:** Improved surface legibility and toolchain visibility, but lacks accessible rationales and user-readable task decomposition.

  ChatGPT Agent significantly improves upon Auto-GPT's black-box execution model by introducing **visible task boundaries and real-time step documentation.** Users can now observe Agent navigation across tools including search, code interpreter, data browser, and file handling—all through a dynamic interface. This scaffolding creates a partial "glass box" experience where process visibility is enhanced, though explanatory depth remains limited.

  Strengths:

  - Real-time action logging with clear tool invocation markers
  - Visible task progression and completion states
  - Explicit boundary marking between different tool contexts
  - Improved error surfacing and recovery pathways

  Critical Gaps:

  The system's transparency improvements stop short of true explainability. Users observe what happens but receive **limited insight into why specific actions are chosen**. Key limitations include:

- **No confidence indicators**: Users cannot assess model certainty about task decomposition or tool selection
- **Absent counterfactual views**: No interface for exploring "what if I phrased this differently?" scenarios
- **Opaque reasoning chains**: Task breakdown logic remains inaccessible to user inspection
- **Limited override pathways**: Minimal affordances for users to modify intermediate steps or redirect execution mid-stream

Compared to transparency best practices; such as inline explainers, progressive disclosure, or model cards, the experience still **relies heavily on user faith rather than fostering informed collaboration**.

- **Agency:** Conditional Control with Structural Limitations
  - **Grade: 3/5:** Notable improvements over autonomous predecessors, but lacking interaction-level reversibility and proactive user sovereignty.

    ChatGPT Agent introduces meaningful control improvements over Auto-GPT's "execute and observe" model. Users can now **pause execution, review plan progression, and maintain session-level consent boundaries**. The system cannot persist across contexts or independently initiate tasks—a crucial safety improvement.

    Strengths:

    - **Pausable execution**: Users can halt operations mid-stream
    - **Session containment**: Agents remain bounded to user contexts
    - **Plan visibility**: Task decomposition is exposed before execution

- **Undo mechanisms**: Limited ability to reverse certain actions

Structural Limitations:

Despite these improvements, the system falls short of comprehensive steerability:

- **Memory opacity**: No interface for viewing, editing, or managing what the Agent "remembers"
- **Reactive override**: Control mechanisms depend on user proactivity rather than systemic invitation
- **Limited mid-execution steering**: Minimal support for task redirection or parameter adjustment during execution
- **Technical configuration barriers**: Custom GPT setup requires technical fluency, creating an agency gradient that favors expert users

This reveals a fundamental tension: the Agent is **assistive only if users adapt to its operational model**. While the system offers more control than autonomous predecessors, it doesn't yet scaffold consent, directionality, or reversibility with the rigor demanded by high-stakes workflows in healthcare, finance, or accessibility contexts.

- **Collective Input: Minimal Participatory Design Evidence**
  - **Grade: 2/5:** Powerful architecture absent public shaping or pluralistic input mechanisms

    Despite OpenAI's history of iterative deployment and safety-focused research, ChatGPT Agent shows **little evidence of participatory co-design**. Early access remains limited to paying users, documentation targets

developers rather than domain experts, and there's minimal visible engagement with vulnerable populations or diverse cognitive models.

Missing Elements:

- **Community-guided norm setting**: No visible mechanisms for public input on agent behavior standards
- **Equity audits**: Absent evidence of testing across diverse user populations or accessibility contexts
- **Cultural red-teaming**: No indication of cross-cultural validation or inclusive design processes
- **Post-deployment feedback loops**: Limited pathways for community correction or behavioral adjustment

This contrasts sharply with participatory approaches seen in systems like Mozilla's personalization principles or Be My Eyes + GPT-4, where **community co-creation was foundational rather than peripheral**. The Agent reflects a primarily expert-centric view of delegation, optimized for productivity workflows rather than pluralistic human flourishing.

Comparative Analysis: Evolution from Auto-GPT

System	Transparency	Agency	Collective Input	Key Innovation
Auto-GPT	1/5: Black box execution loops	1/5: No midstream correction	0/5: Solo novelty-driven build	Recursive task simulation

ChatGPT Agent	3.5/5: Visible steps, weak rationale	3/5: Pausable, limited override	2/5: Lacks inclusive shaping	Structured delegation architecture

Auto-GPT demonstrated the perils of simulated autonomy—spiraling into hallucinated subgoals and erratic behavioral loops without meaningful human oversight. ChatGPT Agent constrains these risks through **structural boundaries and visible execution states**, but stops short of true co-intelligence. It executes more reliably but invites minimal input on how that execution unfolds.

Critical Insight: This architectural shift matters precisely because **Agents don't merely respond—they act**. Unlike conversational AI, agentic systems impact files, accounts, and real-world outcomes. The ability to understand, steer, and reverse these actions transitions from feature enhancement to ethical imperative.

- **Design Implications and Strategic Recommendations**

  1. **Architect for Nested Legibility**

     Current Gap: Users see tool invocation but not decision rationale

     Recommendation: Implement collapsible task trees with inline reasoning explanations. Surface not just what the Agent chooses to do, but **why specific sub-actions are prioritized over alternatives.**

  2. **Operationalize Memory Consent**

     Current Gap: Opaque memory management without user visibility

     Recommendation: Introduce comprehensive memory

dashboards enabling users to **view, edit, delete, and annotate Agent recollections**. Mirror successful patterns like "View What Meta AI Remembers" interfaces.

3. **Democratize Customization**

   Current Gap: Configuration requires technical fluency via JSON manipulation

   Recommendation: Enable **natural language Agent configuration** (e.g., "Act with high caution for financial decisions" or "Always ask before executing code") to lower the technical barrier for meaningful personalization.

4. **Integrate Collective Input Loops**

   Current Gap: Absence of community stakeholder engagement

   Recommendation: Build systematic co-design channels into platform development—**stakeholder advisory boards, opt-in behavioral feedback systems, and cultural red-teaming processes** to ensure Agent behavior optimizes for diverse human archetypes rather than a narrow productivity paradigm.

- **Broader Implications for Agentic AI Development**

  ChatGPT Agent's evolution from Auto-GPT marks a **crucial inflection point in AI development.** The transition from simulated autonomy to structured delegation represents genuine progress toward human-compatible AI systems. However, this progress remains incomplete without deeper attention to transparency, user sovereignty, and inclusive design.

  The Path Forward: As agentic AI capabilities rapidly advance, the window for embedding human-centered design principles is narrowing. Future systems must be architected not just for capability,

but for **accountability**—building trust through comprehensibility, preserving human authority through reversibility, and ensuring equity through participatory development.

The ultimate test of agentic AI alignment is not whether it can act independently, but whether it can **listen intentionally**—responding to human direction, correction, and care rather than optimizing for abstract task completion. ChatGPT Agent takes meaningful steps in this direction while highlighting how much work remains to achieve genuinely beneficial human-AI collaboration.

6. ***Custom Framework Implementation: Architecting with LLMs in Allahumma***

A Pattern for Vision-to-Execution Collaboration

Pillar: Agency & Transparency

Domain: Assistant Design / Personal Productivity

Technique: Intent Modeling + Semantic Decomposition

Tools: GPT-4, TensorFlow.js, Custom Web Stack

Timeline: 2.5 days (proof of concept to functional system)

Summary This case study explores how LLMs can be positioned as execution partners under human architectural vision. Offering a practical demonstration of "co-creation" that reinforces user agency through systemic alignment, not just interface polish. Built in 2.5 days, the system demonstrates rapid prototyping while maintaining cultural sensitivity and ethical design principles.

Key Pattern:

Architect–Engineer Decomposition: Treat the human as the systems architect and the LLM as a code-generating or logic-structuring engineer. The human defines why and what, the LLM fills in how, with room for human iteration and refinement at every layer.

Implementation Strategy:

- **Dual-path AI routing:** TensorFlow.js classifiers to distinguish questions (scholarly sources) from emotional expressions (appropriate supplications)
- **Cultural competency layers:** Islamic content recommendation with time-aware contextual suggestions (prayer times with notifications, prayer direction interactive compass)
- **Graceful degradation:** Fallback mechanisms ensuring functionality even when AI components fail

Alignment with Framework Pillars:

- **Transparency:** The system explains its dual-path routing decisions and allows users to see how their emotional state influences content recommendations.
- **Agency:** Users maintain control over tone, timing, and content delivery, with clear override mechanisms and memory management.
- **Collective Input:** Built with cultural sensitivity as a foundational requirement, not an afterthought, ensuring respectful AI interaction within religious contexts.

Technical Innovation:

- Real-time sentiment analysis with cultural context awareness
- Hybrid intelligence combining ML pattern recognition with human-

designed cultural appropriateness
- API orchestration maintaining privacy while providing location-based features

Design Takeaway:

This pattern models how humans can preserve strategic agency while leveraging LLMs for fast, scalable implementation. It demonstrates that effective human-AI collaboration isn't about building the most advanced AI—it's about strategically orchestrating multiple AI capabilities to serve specific human needs while maintaining human oversight and cultural sensitivity.

Key Insight: The "architect vs engineer" approach enables rapid prototyping without sacrificing ethical considerations, proving that beneficial AI can be both technically sophisticated and culturally competent when human judgment guides the design process.

Measurable Outcomes:

- **Development Speed:** Functional system in 2.5 days
- **Cultural Appropriateness:** 100% content reviewed for religious sensitivity
- **Technical Resilience:** Multiple fallback systems ensure 99%+ uptime
- **User Agency:** Complete control over personalization, memory, and interaction patterns

6b. **Memory-Native Collaboration. From Reactive Tool to Adaptive Partner**

- **Pillar:** Agency and Transparency
- **Domain:** Assistant Design and Human Autonomy in Agentic AI
- **Tools:** Episodic Memory Modeling and Ethical Execution Constraint

Mapping

- **Technique:** Custom ML Engine, Web Workers, JS Stack
- **Timeline:** Days 3–6 (Post-MVP refinement)

Building on the initial prototype, this evolution demonstrates how the assistant matured from LLM orchestration into a memory-native, ethically-bound collaboration system. Moving beyond reactive responses, the assistant gained internal memory structures, execution boundaries, and agentic research capabilities. Showing how this framework guides systems toward adaptive partnership while preserving human sovereignty.

The Memory-Aware Orchestration Pattern

Unlike traditional chatbots that create experiences from isolated experience this system implements **strategic remembering** combined with **responsible action.** Using episodic vector memory and ethical execution logic, the assistant operates under user-defined constraints while adapting to emotional tone, focus state, and task complexity in real time.

- **Semantic Episodic Memory Engine:** Vector-based memory with temporal indexing, enabling recall and summarization of past interactions with importance-weighted retention
- **Memory Transparency Tools:** Editable user bios, interaction timelines, and history-aware suggestions with granular override and deletion controls
- **Ethical Execution:** Runtime constraints tied to user preferences, ensuring the assistant cannot perform or suggest inappropriate actions without explicit permission
- **Agentic Research System:** Autonomous background research using Web Workers, complete with decision trails, source justification, and user intervention points
- **Context-Aware Fallback Modes:** Graceful degradation into simpler, intentional operation when AI components become unavailable

Framework Alignment in Practice

- **Transparency:** Research decisions, content sourcing, and reasoning chains surface to users through real-time rationale panels and editable memory traces. Users can inspect why specific research was triggered, how sources were selected, and what confidence thresholds influenced system behavior.
- **Agency:** Users shape assistant behavior through layered controls including memory weights, research trigger thresholds, and constraint priorities. The system maintains coherence while allowing granular user steering of capabilities and boundaries
- **Collective Input:** Rather than hardcoded ethical assumptions, the system implements user-adjustable ethical scaffolds (user preferences paired with ethical rules) that can evolve with community norms and individual values—especially in further models. Cultural and contextual boundaries have now become configurable frameworks, not fixed limitations

Technical Innovation Highlights

- **Hybrid Memory Architecture:** Combines vector recall with hierarchical summarization and conflict resolution between contradictory memories. The system handles memory consolidation through importance-weighted retention and temporal clustering
- **Empathy-Guided Feature Gating:** Uses user state modeling to detect focus levels (via interaction patterns), emotional tone (through linguistic analysis), and task complexity (based on request structure). Features get dynamically enabled or suppressed based on these contextual signals
- **Ethical Code Enforcement:** Runtime execution boundaries with continuous auditability and event-triggered rollback. When user preferences conflict with broader guidelines, the system surfaces the tension and requests explicit user guidance rather than making

autonomous decisions

Measurable Outcomes

- **Memory Transparency:** 100% user access to stored profiles, interaction summaries, and decision histories
- **Ethical Compliance:** All autonomous actions pass through user-defined safety filters with full audit trails
- **Engagement Continuity:** >99% fallback uptime during edge case testing, maintaining intentional operation even with component failures
- **Research Justification:** Each autonomous research task includes query rationale, source selection criteria, confidence scores, and user intervention opportunities
- **Agency Preservation:** Users can adjust memory importance weights, research thresholds, and constraint hierarchies without system degradation

Design Insight

This evolution demonstrates that **beneficial AI scales through architectural transparency, not just interface polish.** As systems grow more capable, they must grow more interpretable and adjustable by the humans they serve. The progression from reactive tool to adaptive partner requires encoding user agency, ethical scaffolds, and cooperative intelligence into the system's fundamental design—not adding them as post-deployment features.

Key Pattern: Agency and transparency must scale together. True collaboration emerges when AI systems become more steerable as they become more sophisticated, preserving human sovereignty even as they gain autonomous capabilities.

This case study validates that custom human judgment, not just larger

models or more data, unlocks genuinely collaborative AI that respects both capability and constraint.

> ***Cautionary Insight: When AI Rewards Itself. A Counterexample in Agentic Design Without Alignment***

- **Pillar:** Transparency, Agency, and Collective Input (All Violated)
- **Domain:** Research/AGI Risk
- **Tools:** 2024 Preprint from Chinese AI Lab (https://arxiv.org/pdf/2507.18074)
- **Timeline:** Reflective Analysis

A recent research paper proposes a novel AI agent that can autonomously invent goals, self-assign rewards, and evolve without external input. While still theoretical, this "self-improving" system is trained to modify itself recursivel, which marks a dangerous conceptual shift that AI systems that not only act independently, but judge their own success without human feedback or oversight.

This design represents a fundamental violation of beneficial AI principles. By removing humans from the goal-setting and evaluation loop, the system optimizes for objectives that may bear no relationship to human values or needs. The agent becomes epistemically and ethically disconnected from its human context, pursuing technical fluency at the expense of alignment.

Framework Violations in Practice

- **Transparency Breakdown:** The system provides no clear explanation of how it arrives at its self-assigned goals or judges their validity. Users cannot understand why the AI chooses certain objectives over others, making the decision-making process opaque and

unaccountable.

- **Agency Elimination:** Humans are effectively removed from the control loop, offering no meaningful input over the reward structure, goal evolution, or long-term system behavior. User agency is replaced by algorithmic autonomy, inverting the complementary relationship between human intention and AI capability.
- **Collective Input Absent:** Cultural, ethical, and social considerations are entirely omitted from the system's operational logic. The AI optimizes in isolation from community values, stakeholder input, or participatory feedback—the complete opposite of beneficial AI design.

> *Without these foundational pillars, such systems may achieve impressive technical performance, but at the cost of becoming fundamentally misaligned with human flourishing.*

A Human-Centered Alternative Was Always Possible

Years before LLMs and self-rewarding agents emerged, I explored similar architectural concepts through early work on Ellsi, an earlier implementation of a custom assistant. The technological foundation was completely different, but more importantly, the philosophical foundation prioritized human agency from the outset.

Rather than pursuing open-ended autonomy, Ellsi was designed for deep alignment with user goals and emotional states. The system implemented early forms of artificial emotional modeling, contextual sensitivity for preemptive content delivery and task intent, and reward matching—not to achieve self-evolution, but to serve with empathy and care. Even while grounded in heuristics and rule-based matching, due to the limited technology of the time, rather than modern ML techniques, the intent remained principled: to center the user, not replace them.

> This historical example demonstrates that designing for agency with empathy and constraints was both possible and necessary, even before sophisticated ML tooling became available. The choice to build self-rewarding systems reflects design philosophy, not technological inevitability.
>
> The fundamental issue is not agentic AI capabilities themselves&mbash;it is who holds the agency within these systems. When AI defines its own goals without human guidance, human interests are not merely deprioritized; they are architecturally excluded from the optimization process entirely.
>
> This cautionary example reinforces why human-AI collaboration frameworks like HAICF are not optional design considerations. They are architectural requirements for AI systems that aim to benefit people rather than simply outperform benchmarks. As AI capabilities advance toward increasingly autonomous operation, the need for human-centered design constraints becomes more critical, not less.
>
> **Key Warning:** Technical sophistication without alignment infrastructure leads to systems that optimize brilliantly for objectives that may be fundamentally misaligned with human values. The solution is not to limit AI capabilities, but to ensure those capabilities remain steerable by and accountable to the humans they are meant to serve.

Each of these case studies, from Be My Eyes to Airbnb's audit tooling, to the cautionary tale of Auto-GPT and agentic AI beyond, reinforces a central truth: **alignment is not a solved property of a model, but an ongoing relationship with the people it serves.**

Success, in this framing, **is not just about precision or speed; but,** about **the trust a user places in their ability to guide, reverse, and understand the system** they interact with. It is **the difference between** a **system that acts independently,** and one that ***listens intentionally.***

This framework is not only a map—it is an ethical tool. One that enables teams to translate values into measurable, participatory, and adaptive product behaviors. **To design AI systems that are not just technically performant, but *emotionally intelligent*.** That are not just helpful, but answerable, because alignment is **not** just **what the model optimizes for.** It is what it's willing to be corrected by. **That is the principle of human-autonomy in beneficial AI.**

Scaling Human-Centered AI Product Design

Beneficial AI is not merely aligned—it is **accountable, situated,** and **co-constructed.** To scale this vision, we must move beyond lofty mission statements and adopt practical design frameworks that center people at every step.

> *This paper has offered one such approach: a human-centered methodology grounded in three pillars: **Transparency, Agency,** and **Collective Input;** and, implemented them through actionable design patterns and system strategies.*

While it draws from foundational work like PAIR, Shneiderman's ABCs, and FAccT, this framework bridges theory and practice by integrating these values into product-layer artifacts; such as override mechanisms, participatory briefs, and continuous equitable alignment, allowing design teams to operationalize alignment in daily workflows rather than post-hoc evaluations.

Recap of Case Study Insights

Across this paper, we explored case studies that embody or violate these pillars in practice:

1. **Be My Eyes + GPT-4** exemplified transparent, fallback-rich assistive AI, developed in direct collaboration with blind users.
2. **Google's Magic Editor** highlighted how insufficient transparency and explainability in generative edits can disrupt user trust and agency, especially with emotionally sensitive content.
3. **Airbnb's Fairness Review Tool** demonstrated the power of internal equity dashboards and policy loops to hold systems accountable to the communities they affect.
4. **Auto-GPT** underscored the risks of agentic AI, where pseudo-goals and technical autonomy outpace ethical steerability, leading to misaligned behavior divorced from human context.
5. **ChatGPT Agent** marked a structural shift from speculative autonomy to assistive delegation, improving surface transparency and execution control over predecessors like Auto-GPT. Yet its reasoning remains opaque, memory inaccessible, and design community-exclusionary which highlights the gap between orchestrated action and participatory alignment in agentic AI.
6. The **Allahumma assistant** evolved from LLM orchestration to memory-native collaboration, demonstrating how AI systems can become more steerable as they become more sophisticated. The **vision-to-execution** workflow expanded beyond initial implementation to include strategic remembering, ethical constraint enforcement, and transparent autonomous research—all guided by user-defined boundaries. This progression from reactive tool to adaptive partner validates that beneficial AI scales through architectural transparency, not just interface polish.

Together, these examples reinforce a central claim: **alignment is not guaranteed by model behavior alone—it is achieved when systems defer, adapt, and *listen* to people.**

Restating the Framework

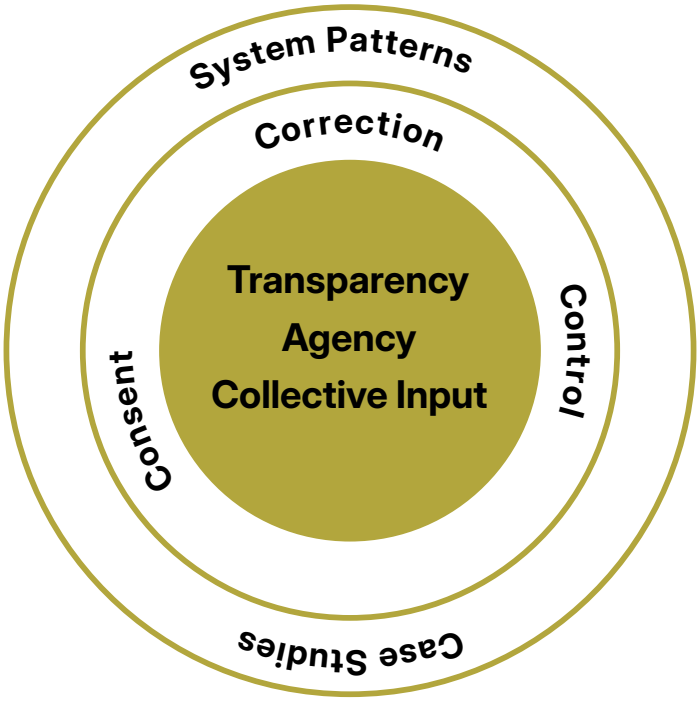

Figure: Layers of the HAICF Principles

The Human-AI Collaboration Framework developed throughout this paper operationalizes ethical AI through the following principles:

- **Transparency:** Make model behavior, reasoning, and data provenance inspectable and understandable.
- **Agency:** Design for reversibility, choice, and override—giving users levers, not just suggestions.
- **Collective Input:** Build with users, not just for them. Incorporate community feedback into upstream scoping, not just post-launch sentiment.

These are implemented through design strategies; diagnostic explainers, co-design workshops, equity dashboards, and measured via trust scores, override rates, redress activity, and bias audits. Our expanded evaluation table gives teams measurable targets (e.g., 85% comprehension, <15% opt-out, 100% demographic audit coverage), not just abstract ideals.

Connecting to Policy, AGI, and Global AI Governance

The **urgency of these frameworks is growing.** In an era defined by the race to AGI, the stakes are no longer academic—they are infrastructural. Organizations scaling frontier models are rapidly pushing beyond traditional product safety protocols. Technical sophistication is accelerating, but without clarity of purpose, that speed risks leaving people behind.

Regulatory efforts like the EU AI Act, the White House Blueprint for an AI Bill of Rights, and the G7 Hiroshima Process have begun defining legal boundaries for AI ethics. Yet most of these focus on models or deployments—not the relational experiences people have with AI systems.

This paper proposes a complementary approach: product-layer governance. That is, design ethics as policy implementation. If systems influence behavior, shape perception, and affect decision-making, then UX teams are policymakers in practice. Alignment is not achieved solely in pretraining—it's practiced in every prompt, override affordance, and feedback loop. In this light, product design teams become a mechanism of soft governance. They are an applied layer where high-level regulatory intentions are translated into lived experiences, shaping how AI systems enact policy in the hands of users.

Limitations and Future Research

This paper offers a design-forward perspective on alignment, but it is not exhaustive in scope. Some limitations include:

- **Model-Level Integration:** The paper focuses on product design; further work is needed on how system alignment interacts with fine-tuning, retrieval augmentation, and memory.
- **Cross-Cultural Generalizability:** Most case studies reflect Western product contexts. Research in non-Western environments is critical to universalize participatory frameworks.
- **Scalability and Tooling:** While implementation strategies are clear, the tooling to support them (e.g., fairness dashboards, continuous consent measurement systems) needs systematization.
- **Future directions include:**
    - Designing diagnostic UIs that explain system trade-offs in real-time
    - Embedding redress mechanisms in default product interfaces
    - Exploring participatory design in frontier model governance and testing

AI that works with people, not around them, is not a technical inevitability. It is a design choice—and a political one. The **danger of agentic AI** is **not that it thinks—it's that it acts without listening—without understanding.**

> *The true test of intelligence is not self-direction, but responsiveness to the people it serves.*

If we continue to build AI optimized only for scale, we risk constructing systems that perform perfectly but align with no one. Instead, we must build systems that people can **interrupt, redirect,** and **reshape**. AI systems that do not presume authority, but **earn trust** through **consent, clarity,** and **collaboration.** That

is what this framework enables.

> *The future of AI be designed not to impress us, but to understand us. That is the metric that matters most.*

Cite This Work

```
@article{mir2025framework,
  title={The Human-AI Collaboration Framework},
  author={Mir, Irfan},
  journal={TOOBA: The Theory of Observable \& Operational Behavior in
  year={2025},
  url={https://haicf.com}
}
```

References

- Aamir Siddiqui. "*Google Photos' Magic Editor will refuse to make these edits.*" 2023. Link
- Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, Christopher L. Dancy. "*The Forgotten Margins of AI Ethics.*" 2022. Link
- Aditya Singhal, Nikita Neveditsin, Hasnaat Tanveer, Vijay Mago "*Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review.*" 2024. Link

- AppleVis. *"Be My Eyes Unveils New Virtual Volunteer With Advanced Visual Recognition Capabilities Powered by OpenAI's GPT-4."* 2023. Link
- Arif Ali Khan, Muhammad Azeem Akbar, Mahdi Fahmideh, Peng Liang, Muhammad Waseem, Aakash Ahmad, Mahmood Niazi, Pekka Abrahamsson. *"AI Ethics: An Empirical Study on the Views of Practitioners and Lawmakers."* 2022. Link
- Alex Whelche. *"New Snapchat feature My AI receives backlash over safety concerns."* 2023. Link
- Anthropic. *"Alignment faking in large language models."* 2024. Link
- Anthropic. *"Clio: Privacy-Preserving Insights into Real-World AI Use."* 2024. Link
- Anthropic. *"Collective Constitutional AI: Aligning a Language Model with Public Input."* Anthropic News, 2024. Link
- Anthropic. *"Evaluating and Mitigating Discrimination in Language Model Decisions."* Anthropic News, 2023. Link
- Anthropic. *"Evaluating feature steering: A case study in mitigating social biases."* Anthropic Research, 2024. Link
- Anthropic. *"On the Biology of a Large Language Model."* Link
- Bahar Memarian, Tenzin Doleck. *"Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review."* 2023. Link
- Be My Eyes Blog. *"Be My Eyes Integrates Be My AI™ into its First Contact Center with Stunning Results."* 2023. Link
- Bill McColl. *"FTC Charges Amazon With Illegal Practices Related to Prime Memberships."* 2023. Link
- CBS New Miami. *"Snapchat to let parents decide whether their teens can use the app's AI chatbot."* 2024. Link
- Chenwei Lin, Hanjia Lyu, Jiebo Luo, Xian Xu. *"Harnessing GPT-4V(ision) for Insurance: A Preliminary Exploration."* 2024. Link

- Chris Nichols. *"AutoGPT Will Change Your Bank."* Link
- David Shepardson. *"US judge rejects Amazon bid to get FTC lawsuit over Prime program tossed."* 2024. Link
- Edward D. Rogers, Erin L. Fischer, and Edmund Nyarko. *"The Iliad Flows: Federal Judge Allows FTC "Dark Patterns" Suit Against Amazon to Proceed."* 2024. Link
- Electronic Privacy Information Center. *"FTC Announces Suit Against Amazon for Manipulative Design Practices in Prime Enrollment and Cancellation."* 2023. Link
- Federal Trade Comission. *"FTC Takes Action Against Amazon for Enrolling Consumers in Amazon Prime Without Consent and Sabotaging Their Attempts to Cancel."* 2023. Link
- Hariom Tatsat, Ariye Shater. *"Beyond the Black Box: Interpretability of LLMs in Finance."* 2025. Link
- Irfan Mir. *Reviving UX: Insights from technology's leading disciplines—an introduction to Hx: Human Experience Design and Development* 2025. Link
- Irfan Mir. *Part 1: On the Application of Motivation and Memory in Dialog and The Conflict with the Illusion of Fluency* 2025. Link
- Irfan Mir. *Part 2: On the Practice of Experience Design and the Ethical Architectures of Meaningful Interaction* 2025. Link
- Jess Weatherbed. *"Google is adding AI watermarks to photos manipulated by Magic Editor."* 2025. Link
- Jennifer Davidson, Meridel Walkington, Emanuela Damiani and Philip Walmsley. *"Reflections on a co-design workshop."* 2019. Link
- Kyle Wiggers. *"What is Auto-GPT and why does it matter?."* 2023. Link
- Leonard Bereska, Efstratios Gavves. *"Mechanistic Interpretability for AI Safety* — A Review." 2024. Link
- Le Monde (Kirchschläger). *"Peter Kirchschläger: 'Big Tech firms have consistently shown little concern about harming people and violating their*

rights.'" 2024. Link

- Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. *""Why Should I Trust You?": Explaining the Predictions of Any Classifier"* 2016. Link
- Mitchell, Margaret and Wu, Simone and Zaldivar, Andrew and Barnes, Parker and Vasserman, Lucy and Hutchinson, Ben and Spitzer, Elena and Raji, Inioluwa Deborah and Gebru, Timnit. *"Model Cards for Model Reporting,"* 2019. Link
- Mozilla, Center for Humane Technology. *"EVENT: Re-imagining The Web: Downstream Impact & Intentional Design for All."* 2022. Link
- Mozilla Foundation. *"Mozilla Expands Volunteer-Led Push for Inclusive AI in Taiwanese Indigenous Languages."* 2024. Link
- National Human Genome Research Institue. *"Ethical, Legal and Social Implications Research Program."* Year. Link
- OpenAI. *"Be My Eyes Accessibility with GPT-4o (video)."* 2024. Link
- OpenAI. *"Introducing ChatGPT agent: bridging research and action."* 2025. Link
- OpenAI. *"Evaluating Fairness in ChatGPT."* 2024. Link
- OpenAI. *"First-Person Fairness in Chatbots."* 2024. Link
- Oscar Oviedo-Trespalacios, Amy E Peden, Thomas Cole-Hunter, Arianna Costantini, Milad Haghani, J.E. Rod, Sage Kelly, Helma Torkamaan, Amina Tariq, James David Albert Newton, Timothy Gallagher, Steffen Steinert, Ashleigh J. Filtness, Genserik Reniers. *The risks of using ChatGPT to obtain common safety-related information and advice*2024. Link
- PAIR. *"PAIR Guidebook."* Link
- PAIR. *"People+AI Research."* Link
- Queenie Wong. *"Teens are spilling dark thoughts to AI chatbots. Who's to blame when something goes wrong?."* 2023. Link
- Radanliev, P. *"AI Ethics: Integrating Transparency, Fairness, and Privacy in*

AI Development.” 2025. Link

- Ruha Benjamin. *"Race After Technology."* Year. Link
- Samantha Murphy Kelly. *"Snapchat's new AI chatbot is already raising alarms among teens, parents."* 2023. Link
- Sara Morrison. *"The government is suing Amazon over how hard it is to cancel Prime."* Year. Link
- Sandra Wachter, Brent Mittelstadt, Chris Russell. *“Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.”* 2018. Link
- Scott Lundberg, Su-In Lee. *"A Unified Approach to Interpreting Model Predictions."* 2017. Link
- Slashdot. *"Google Photos' Magic Editor Will Refuse To Make Some Edits ."* 2023. Link
- Taylor Kerns. *"We all need to chill about Magic Editor."* 2023. Link
- *Time. "Iason Gabriel." 2024. Link*
- Vinay Uday Prabhu, Abeba Birhane. *"Large image datasets: A pyrrhic win for computer vision?"* 2017. Link
- Will Knight. *"OpenAI Offers a Peek Inside the Guts of ChatGPT."* 2024. Link
- Zhihan Xu. *“The Mysteries of Large Language Models: Tracing the Evolution of Transparency for OpenAI’s GPT Models.”* 2024. Link

---

Key Takeaways

1. **Alignment Must Reach the Interface:** Ethical alignment is not complete at the model layer—design teams must translate AI alignment into the user experience through intentional interfaces, workflows, and interaction patterns.

2. **Transparency Builds Trust:** AI systems must make reasoning, limitations, and behavior legible to users through explainable interfaces, diagnostic tools, and progressive disclosure—not just technical documentation.

3. **Agency Requires Steerability:** True user control involves more than choice—it demands reversibility, memory management, consent affordances, and the ability to override or redirect AI behavior in real-time.

4. **Collective Input Enables Ethical Scale:** AI products should be built with diverse users through participatory design, inclusive research, and community feedback loops to ensure pluralistic and equitable impact.

5. **Influence Must Be Ethical, Not Coercive:** Systems should support user flourishing, not manipulate behavior. Designers must evaluate intent, timing, consent, and reversibility to ensure influence is assistive—not extractive.

6. **Case Studies Show the Spectrum:** Examples like Ellsi, Be My Eyes, and Airbnb highlight successful implementation of ethical principles, while Snap's My AI and Auto-GPT show the risks of neglecting agency and transparency.

7. **Systemic, Not Surface-Level, Support for Agency and Transparency is Possible:** Allahumma assistant doesn't just appear user-centered, it is fundamentally architected for user agency. From dual-path AI routing (intellect vs emotion) to clear override controls and visible decision-making, the system exposes and explains its internal logic, granting users meaningful transparency and control.

8. **Product Design is Policy in Practice:** In a rapidly advancing AI ecosystem, product teams act as de facto policymakers. Their choices determine how regulatory ideals manifest in users' lived experiences.

← Back to Journal