Intent and Boundaries: A Framework for Digital Agency

Daniel Hardman Independent Researcher
daniel.hardman@gmail.com

December 20, 2025

Abstract

In the physical world, human intent is signaled through a rich context of body language, tone, and environment. In the digital realm, however, complex purposes are often compressed into low-fidelity signals, such as a single mouse click. This reduction creates a "semantic void" where users and systems frequently misalign, leading to manipulation, error, and eroded trust. This paper defines a rigorous model of "intent" and introduces the concept of **intent boundaries**—specific thresholds where an external observer's knowledge of an actor's purpose becomes inadequate. By synthesizing principles from bioethics, psychology, and design theory, we propose a framework for recognizing and respecting these boundaries, ensuring that future digital architectures—particularly those driving agentic AI—preserve human agency rather than subvert it.

Keywords: Intent, Agentic AI, Human-Computer Interaction (HCI), Digital Agency, Ethics, Dark Patterns, User Experience (UX), Identity

---

1. The bandwidth problem

Human agency is robust in the physical world because our signaling bandwidth is high. If I walk toward a door, grasp the handle, and turn it, my intent to exit is obvious to anyone watching. The context (walking), the mechanics (turning), and the consequence (opening) are aligned and transparent.

Digital agency is fragile because the bandwidth is low.

Consider a user—we'll call him Arthur—navigating a streaming service on his smart television. He sees a drama he likes and clicks a button labeled "Watch." To Arthur, the intent is sensory and immediate: he wants to see the show. To the service provider, however, that specific click is interpreted as a signal to upgrade Arthur's subscription, alter his billing terms, and execute a recurring payment.

In the physical world, signing a contract looks nothing like opening a door. In the digital world, they can look exactly the same: a single tap of a finger.

This ambiguity creates a dangerous gap. The system assumes a depth of intent that the user's simple action did not actually carry. When Arthur later discovers he has "agreed" to terms he never contemplated, he feels manipulated—and rightly so. The interface relied on a "dark pattern" [1], benefiting the provider by exploiting the low fidelity of the signal.

As we move toward an economy of AI agents, where software will act on our behalf with increasing autonomy, this fragility becomes a security crisis. If an AI cannot distinguish between a user's idle curiosity and their commitment to a contract, it cannot safely represent them. We need a more rigorous way to map the terrain of human purpose.

2. A mental model of intent

To solve this, we first need a definition of intent that is precise enough for systems design. Drawing on the philosophical work of Anscombe [2], we can define intent not merely as a wish, but as a specific cognitive posture.

> **Intent** is a mental stance that explains a choice of action as contributing to a specific purpose.

This definition implies four requirements:

1. **Purpose.** If Arthur clicks "just because," without a goal, he acts without intent.
2. **Action.** Intent must drive a choice. If the action is only contemplated, the intent is merely potential; if no action is imagined, it is an idle wish.
3. **Choice.** A fumble of the remote or a reflex reaction is not intent.
4. **Internal Explanation.** The actor must possess an internal narrative that links the action to the purpose.

Crucially, this internal explanation is often **conditional**. Arthur's intent to watch may carry implicit constraints: "I intend to watch, *provided* the cost is zero."

Furthermore, intent is hierarchical. A single action often supports multiple intents depending on the "horizon" of the purpose. When Arthur clicks "Watch," he holds a **proximate intent** ("I intend to click this button") and an **ultimate intent** ("I intend to relax").

While Arthur is the sole authority on his internal state, he is still bound by logic. He cannot intend to do what he cannot conceive. He implies consequences he can foresee—if he intends to step into the rain, he also intends to get wet—but he does not intend consequences he cannot see. This visibility gap is where the trouble starts.

3. Intent boundaries

Systems fail when they assume they know more than they do. We can formalize this danger zone.

> An **intent boundary** is a threshold where an external party's knowledge of an actor's intent becomes inadequate to justify further action.

It is vital to understand that this boundary is **relative**. It is not a fixed line in the code; it is a function of the observer's ignorance.

- **From Arthur's perspective**, there is no boundary. He knows exactly what he wants (to watch the show). He assumes the cost is zero.
- **From the System's perspective**, there is a massive boundary. The system knows the cost is $50. It does not know if Arthur knows this. Therefore, the system's knowledge of Arthur's intent is "inadequate" to justify the charge.

When the system charges him anyway, it violates the boundary. It substitutes a guess ("he clicked, so he must want to pay") for knowledge.

4. Principles for design

Recognizing these boundaries allows us to move from vague "user friendliness" to rigorous ethical design. We can derive four normative principles.

Principle 1: Recognize the boundary

The first failure of the streaming interface was a failure of recognition. The designers pretended that a button labeled "Watch" provided adequate signal to execute a financial upgrade. They ignored the gap between the user's likely knowledge and the system's action.

Recognizing a boundary is the architectural equivalent of the bioethical standard of **informed consent** [3]. In medicine, a patient's nod is not sufficient legal grounds for surgery because the patient may not understand the risks. Similarly, in software, a click is not sufficient grounds for a high-stakes transaction if the context is ambiguous. We must proactively identify where our knowledge of the user runs out.

Principle 2: Move the boundary

If a boundary exists, we do not always have to stop the user with a pop-up dialog. Good design can often **move** the boundary to a place where friction is lower.

Consider the "Watch" button again. If the button were colored red and labeled "Rent for $5.99," the ambiguity would vanish.

This touches on what Norman calls **affordances** and **signifiers** [4]. By changing the signifier (the label and color), we change the user's understanding *before* they act. The intent boundary—the point of "inadequate knowledge"—shifts. When the user clicks the red button, the system now *knows* that the user understands the cost. The action is now safe.

Principle 3: Peek across the boundary

When certainty is impossible, systems can sometimes "peek" across a boundary by lowering the stakes.

If the streaming service cannot change the button design, it could start playing the video with a superimposed message: "Premium Content. Cancel within 60 seconds to avoid charges."

This approach allows the user's proximate intent (watching) to proceed while provisionally verifying their ultimate intent (paying). It respects the boundary by making the crossing reversible and low-risk.

Principle 4: Never sneak across

The defining characteristic of "dark patterns" [1] is the surreptitious crossing of intent boundaries. Sneaking takes many forms:

- Bundling a legitimate intent (watching) with an undisclosed outcome (subscribing).
- Burying data reuse permissions in a forty-page terms of service document.
- Logging user actions inaccurately ("User chose to upgrade" rather than "User clicked Watch").

This principle aligns with Cavoukian's concept of **Privacy by Design**, specifically the requirement for "Respect for User Privacy" [5]. An agent that sneaks across boundaries undermines the very concept of agency. It treats the user not as a principal to be served, but as a resource to be mined.

5. Conclusion

Mishandled intent boundaries are at the heart of the modern trust crisis in technology. As we build the next generation of digital agents, we must do better. We cannot treat intent as a binary switch that is either "on" or "off." We must build architectures that recognize the semantic void between a click and a commitment.

The best systems will be those that see intent boundaries not as obstacles to conversion, but as the necessary borders of human autonomy.

References

[1] Brignull, H. 2010. Dark Patterns: Deception vs. Honesty in UI Design. *A List Apart.*

[2] Anscombe, G.E.M. 1957. *Intention.* Harvard University Press.

[3] Faden, R.R. and Beauchamp, T.L. 1986. *A History and Theory of Informed Consent.* Oxford University Press.

[4] Norman, D.A. 2013. *The Design of Everyday Things.* Basic Books.

[5] Cavoukian, A. 2009. *Privacy by Design: The 7 Foundational Principles.* Information and Privacy Commissioner of Ontario.