universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Designing Understanding for AI Interfaces
From Communication Models to Interaction Design and Adaptability

verfasst von | submitted by

Elisabet Delgado Mas

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 013
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Joint-Degree Middle European interdisc. master prog. in Cognitive Science
Betreut von | Supervisor:	Martin Takac

Table of Contents

This page is
intentionally left blank.

Abstract

This thesis explores human and artificial intelligence (AI) interaction through the lens of interface design and communication theory. As AI increasingly augments human activity, the primary bottleneck for positive impact gradually changes from technical optimization to human cognition and the interaction patterns we establish with AI. We argue that the interfaces of conversational agents (CAs), such as ChatGPT and Claude, often fail to communicate invisible capabilities, limitations, and uncertainty, all compounded by how differently signifiers are interpreted across people and cultures, which we define as the four challenges of AI signification. By synthesizing concepts like Norman's design principles and Shannon-Weaver's and Schramm's communication models, we analyze where and how meaning is lost in communication, and propose moving from static design toward human-centered adaptive interfaces. To translate diagnosis into intervention, we introduce dynamic signifiers, interface elements that evolve with the CA and the user, and present three implementable designs for current CA interfaces: (1) contextual warning feedback that does not interrupt conversational flow, (2) a capability tracking wish list that transforms moments of system failure into documented opportunities for user engagement and development feedback, (3) a cognitive changelog that makes CA evolution transparent and personally relevant through an adaptive dashboard and pedagogical metaphors. Integrating cognitive science, communication theory, and design practice, this research offers both the theoretical vocabulary and practical starting points for making AI interfaces more human-centered, fostering discoverability of affordances and useful conceptual models of AI.

Zusammenfassung
Abstract in German

Diese Arbeit untersucht die Interaktion zwischen Mensch und künstlicher Intelligenz (KI) aus der Perspektive von Interface-Design und Kommunikationstheorie. Je mehr KI menschliche Aktivitäten erweitert, desto stärker verlagert sich der Engpass für positive Wirkung: weg von technischer Optimierung, hin zu menschlicher Kognition und den Interaktionsmustern, die wir mit KI etablieren. Wir argumentieren, dass die Interfaces von Conversational Agents (CAs), wie ChatGPT und Claude, häufig daran scheitern, unsichtbare Fähigkeiten, Limitierungen und Unsicherheit zu kommunizieren. Erschwert wird dies zusätzlich dadurch, wie unterschiedlich Signifikanten von verschiedenen Menschen und Kulturen interpretiert werden, was wir als die vier Herausforderungen der KI-Signifikation definieren. Durch die Synthese von Konzepten wie Normans Designprinzipien und den Kommunikationsmodellen von Shannon-Weaver und Schramm analysieren wir, wo und wie Bedeutung in der Kommunikation verloren geht, und schlagen einen Übergang von statischem Design hin zu menschzentrierten adaptiven Interfaces vor. Um Diagnose in Intervention zu übersetzen, führen wir dynamische Signifikanten ein, Interface-Elemente, die sich mit dem CA und dem Nutzer weiterentwickeln, und präsentieren drei umsetzbare Designs für aktuelle CA-Interfaces: (1) kontextuelles Warn-Feedback, das den Gesprächsfluss nicht unterbricht, (2) eine Capability-Tracking-Wunschliste, die Momente des Systemversagens in dokumentierte Möglichkeiten für Nutzerengagement und Entwicklungsfeedback verwandelt, (3) ein kognitives Changelog, das die Evolution von CAs durch ein adaptives Dashboard und pädagogische Metaphern transparent und persönlich relevant macht. Durch die Integration von Kognitionswissenschaft, Kommunikationstheorie und Designpraxis bietet diese Masterarbeit sowohl das theoretische Vokabular als auch praktische Ansatzpunkte, um KI-Interfaces menschzentrierter zu gestalten und die Auffindbarkeit von Affordanzen sowie nützliche konzeptuelle Modelle von KI zu fördern.

"The saddest aspect of life right now
is that science gathers knowledge
faster than society gathers wisdom"

Isaac Asimov, 1988

This even page is
intentionally left blank.

1. Introduction

This thesis aims to address a rather neglected side in AI development: the design of AI interfaces, and to what extent they shape the user's conceptual models of these technologies. We argue that the interface design represents the weakest link in the chain that transforms AI into a practical tool for its users and a valuable product for society. And we propose that we can convey better conceptual models to the users through it. The expected impact of this thesis lies in its potential to reshape how explainable AI is designed and implemented. What motivates this work is the recognition that **even subtle strategic tweaks in interface design can save millions of hours worldwide** by making AI interactions more efficient and intuitive. The scale of impact is enormous; with AI increasingly enhancing and complementing human cognition and activities, understanding and optimizing this interaction can shape AI development for decades to come.

AI is a tool with the potential to augment human cognition, and positively transform industries and communities. However, its impact hinges on its usability for us(ers), humans. This thesis aims to show that AI interfaces and explanations empower rather than frustrate us(ers), by looking at ways to improve AI interfaces and explanations, applying human-centered design principles, checking if they are intuitive, accessible, and aligned with how people think and communicate. By enhancing user interaction with AI, we aim to bridge the gap between complex technology and everyday users.

The AI industry has been largely ruled by technical disciplines like computer science, mathematics, and engineering. These prioritize goals such as optimization, efficiency, and accuracy. But often human interaction is neither of these things. Therefore, communication usually becomes the weakest point of AI-driven systems. At the core of all of this rapid technological development and these massive models, lies the simple fact that we are building AI for people. Most of its developers acknowledge that conversational agents (CAs) could be more useful to their clients than they usually are, if understood well and constantly and consistently steered in the right direction. The Asimov quote we started with, encapsulates the result of this trend of having increasingly powerful tools that require far too much effort to learn for the common user, and whose interfaces remain rather opaque.

Bad interfaces lose users to the competition: a feature they deem necessary implemented elsewhere, a functionality they could not find, or frustrating uncertainty can

be enough to abandon a product. The Internet can be a fast-paced world, in which many users have discarded pages and services impulsively in less than a few seconds. Based on assumptions to prioritize our time, people will take what they see as what there is. On top of that, most users do not read manuals or documentation, instead judging entirely by their previous experience and what is directly in front of them: an interface and the system's response to their often poorly formulated queries stemming from faulty conceptual models about AI. As CAs accumulate knowledge, human cognitive constraints and interaction patterns become the limiting factors in transforming it to wisdom. The better AI becomes, the more the bottleneck for positive impact will gradually lie in the interaction patterns we establish with it, and in human cognition. Cognitive science and interaction design offer empirically grounded insights to identify the problem, articulate it, and address these constraints.

For the vast majority of CAs users, the interface is not just important; it is everything they have. Without API access or technical knowledge, the interface becomes their only gateway to interact with the AI behind it, which might determine whether CAs are intuitive or alienating, empowering or inaccessible. Despite playing such a significant role in how people interact with it, interface design and the user's conceptual model of AI remain rather underexplored domains in AI research both in the industry and academia. The absence of forums or sections dedicated to product design, interfaces and conceptual models in the online spaces that are supported and promoted by the industry (from dedicated research sections on their webpage to their online forums or social media channels) underscores this neglect, so does the scarcity of scientific publications and PhD programs in design in general. This thesis aims to emphasize the need for a more integrative approach to AI research and development (R&D), aiming to focus on CAs interfaces and if they can serve as a smooth augmentation of human cognition rather than a barrier to it.

Our approach is necessarily **interdisciplinary**, rooted in the assumption that complex problems, in our case human-AI interaction, rarely stay within the boundaries of a single research field. As David Marr noted, "trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: it just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then does the structure of feathers and the different shapes of birds' wings make sense" [0]. In this thesis, by analogy, the *aerodynamics* of human-AI interaction is where cognitive science, interaction design, and communication theory

come together. We treat our inquiry as a Rubik's Cube: looking at only one face may give the illusion of a solved problem, but it is only by turning the cube that we see how each side constrains and enables the rest, and further making a move on one face of the cube affects the positions of others.

To state it more explicitly, the **overarching** subject of this thesis is human-AI interaction. However, this is an extremely broad field spanning from industrial robots to medical diagnostic systems, encompassing a wide range of applications and contexts. Our **object of study** throughout this thesis is conversational agents (CAs), the artificial systems we engage with on platforms such as ChatGPT, Claude.ai, Mistral or DeepSeek. An agent is a person or thing that takes an active role or produces a specified effect. CAs are conversational because they are driven by large language models (LLMs), fine-tuned for human interaction. Their purpose is to engage users in conversation through their interfaces.

Naming them CAs is deliberate, motivated by both theoretical precision and practical clarity. Within this thesis, we consider **Artificial Intelligence (AI)** to be the capability, skill or attribute of intelligence in an artificial agent, and use **Conversational Agent (CA)** to refer to the actual system possessing it. In the AI community, terms like "model" are more common, but this engineering perspective does not fully capture what users interact with, which may integrate multiple models selectable in a drop-down menu, and even a Mixture of Experts (MoE) architecture behind a single model.

From a research perspective, this presents a unique layering of communicative systems precisely because they operate at two simultaneous levels of mediation: they employ language (a symbolic system) while simultaneously requiring interface design (a visual system, in which verbal language is used and displayed) to make their capabilities accessible. This double mediation creates distinctive challenges that make CAs particularly worthy of investigation.

To place our **focus** on a more specific intersection, we want to examine the graphic user interfaces (GUIs), including the language labels displayed in them, through which humans engage with conversational AI. Why are we focusing primarily on GUIs, rather than other user interfaces? GUIs are still our preferred way to interact with the virtual world, which is where AI primarily dwells. Given this master's thesis has to be physically printed and rendered as a PDF, they also remain more practical to thoroughly research than other kinds such as voice user interfaces (VUIs). Other

formats of narrow interfaces such as command-line interfaces (CLIs) are rather niche, restricted to those comfortable enough with code and typing commands on a terminal. Forms of broad interfaces, such as robots, remain incredibly expensive due to manufacturing and research costs. Conversely, GUIs are accessible, broadly and for free, which makes researching them all the more consequential.

This choice is also an **ethically-motivated decision**. We want to research something both accessible and relevant for the most people possible, and human-AI interaction, especially CAs and their GUIs are becoming increasingly prevalent in many societies. A single minor improvement in a GUI might benefit millions of users. Improving user experience, avoiding user confusion, or even saving a single second of hesitation during navigation matters, but let us see it with a quantifiable example to grasp the dimensions of what we are talking about: for a CA with 500 million weekly users (as, for example, ChatGPT), a single second of improved efficiency in their interactions saves the collective human time from wasting 15.85 years. That is half of the lifetime of the author of these lines, every week.

These interfaces for CAs represent relatively recent communicative artifacts with rapidly evolving patterns, offering particularly fascinating ground for analysis. Few domains are so fast-paced, and its accelerated evolution makes systematic research both intellectually challenging and all the more urgent.

This focus necessarily **excludes** detailed examination of other important interface types and AI applications. This does not diminish the importance of questions regarding more embodied interfaces, industrial robots, specialized AI in medical or scientific domains, and emerging modalities like brain-computer interfaces. They all merit their own dedicated analyses.

Finally, our **approach** aims to be intentionally philosophical, questioning assumptions rather than advocating for particular positions. By integrating perspectives from cognitive science, communication theory, and design practice, this interdisciplinary orientation seeks to develop analytical depth that can inform both theoretical ideation, practical creation, and effective communication to the reader: to learn and communicate.

The **scope** of this thesis consists of two perspectives, but only the first one is developed in this thesis. The first one focuses on what we see: communication, interfaces and user-centered design. The second focuses on what is behind: understanding

AI and interpretability. This first perspective establishes some foundation for the second, ensuring that we fully understand the range of visual cues, feedback and limitations users experience when interacting with CAs like ChatGPT and Claude. It comprises three parts to explore the what, why and how of interface design.

The rest of this thesis is **structured** as follows:

This chapter provides the motivation behind our research focus, connecting industry, research, cognitive science and design. The upcoming chapters are structured in a way to approach our focus through three main clusters of questions: chapter 2 addresses the question of what, chapter 3 the question of why, and chapter 4 the question of how..

Chapter 2 focuses on the definitions of an interface. This chapter provides both a narrow and a broad definition of an interface. While we take the narrow definition as the primary focus for the rest of this thesis, most of the concepts explored are still very valid for the broader concept of interfaces. For instance, the physical instantiation of a robot can serve as an interface between an artificial agent and the world. We argue that these interfaces, whether virtual or physical, must be designed with the same level of care as virtual interfaces. Although some of the concepts we explore and extend in the following two chapters are specific to GUIs, since that is one of the main faces of CAs today, most of them are useful for both virtual and physical interfaces, and even for products or elements of our everyday life.

Chapter 3 focuses on why communication between humans and CAs breaks down, and where. Through Shannon-Weaver's model we trace how signals become messages and where noise distorts them, extending their ideas to account for semantic, effectiveness, and conceptual model interferences. Through Schramm's model we explore what makes communication possible in the first place: overlapping fields of experience, and feedback loops. The section concludes by framing AI interfaces as communicative artifacts.

Chapter 4 looks at the know-how of design from a cognitive science lens to think about interfaces and human-centered design. In this chapter, we start by providing a solid theoretical ground in interaction design which begins with Don Norman's principles. In this context, affordances determine how the object could possibly be used, while signifiers provide clear visual cues about functionality. Mapping ensures that controls correspond logically to their effects, and feedback allows users to see

the consequences of their actions. Conceptual models shape user expectations of how a system works, and constraints help prevent errors by limiting unnecessary or confusing actions.

Finally, **chapter 5** builds on everything elaborated in chapters 2-4 to propose dynamic signifiers (5.1), interface elements that can evolve with the CA and communicate the changing system states and the extent of invisible capabilities. This is exemplified by three design proposals (5.2) that could be implemented in the interfaces of CAs in order to better communicate limitations, uncertainty and the ongoing development of CA capabilities: warning light indicators for messages, a wish list for capability tracking, and a cognitive changelog, and finally adaptive interfaces (5.3).

Chapter 6 concludes the whole thesis.

⚑ What are we researching?

Overarching field: **human-AI interaction**
Object of study: **conversational agents (CAs)**
Focus: **graphic user interfaces (GUIs)** as communicative and designable artifacts
Approach: **interdisciplinary, integrative, inquiring**

2. *What.* Of Interfaces, Robots and Bodies

Imagine if your only way to communicate with another person was through a narrow tube that filtered everything you said and everything you heard. Sound grows cloudy, emotions lose their nuances, questions their playfulness, and jokes arrive a beat too late. This is the fundamental challenge of AI interfaces today. They are constrained channels between two primarily different cognitive systems. While we are rapidly improving what can flow through these channels, less attention has been paid to the characteristics of the channels themselves. Yet these interfaces do not just transmit information; they transform it, shape it, and ultimately determine what is possible in the interaction.

For a more formal definition, the International Organization for Standardization (ISO) describes a user interface as "all **components** of an interactive system (software or hardware) that **provide information and controls for the user** to accomplish specific tasks with the interactive system" (ISO 9241-110:2006) [1]. In our everyday life, we face a vast range of physical interfaces, like the bathroom faucets that control our access to water, the steering wheel that allows us to translate our intentions to a car's mechanical systems, or the elevator buttons (floor numbers, open/close, and up/down call buttons) that let us select our vertical destination. All of these provide information about how to steer or control an artificially built system. When we talk about interfaces in the digital realm, and particularly in the context of AI, we are discussing the virtual analogues of these meeting points between human cognition and artificial systems.

We can also adapt the common understanding of a user interface (and by the common understanding of a user interface, we take "the space where interactions between humans and machines occur" [2]) to our specific context. As educational resources from CalArts aptly describe, the interface is not merely a window but "a bridge between the user and the content that shapes the way the user **experiences the content**" [3]. Interfaces are a concept so ingrained in our daily lives that is often as invisible to us as the air we breathe. When we press a button displayed on the glass screen of our smartphones to check the weather before going out, that touch becomes the bridge between our biological intelligence and silicon-based computation. In turn, to provide us with today's temperature, this silicon-based computation might be interfaced with networks of physical sensors, such as thermometers, and meteorological systems, creating a **chain of interfaces** that extends from our finger-

tip through digital systems and back to the physical world.

> ⚑ **General interface definition**
>
> The **interface** is the space where interactions (between humans and artificial systems) take place. It consists of all the **elements** of an interactive system and their **arrangement**, which provide information and control to the user to perform specific tasks with the interactive system.

2.1 Narrow View, Dimensions of Interface Expansion and Broad View

As the general definition of interface is rather generic, in this sub-chapter, we will look into two views of the concept of an interface with specific examples: **(1)** narrow or conventional and **(2)** broad or embodied. After we introduce the narrow definition, we will suggest three dimensions of possible expansion of artificial interfaces and then connect this discussion with artificial embodiment and thus, the broad view of interfaces.

Traditionally, interfaces have been understood rather **(1)** narrowly (conventionally) as the point of contact between a human and an artificial system. This perspective gives us familiar examples like:

- Graphical user interfaces (GUIs) with the buttons, menus, and other visual elements on our computer desktops or a web page.
- Voice user interfaces (VUIs) enabling spoken interaction with systems like Siri or Gemini.
- Command-line interfaces (CLIs) to type commands on a terminal with an operating system.
- Tangible interfaces that employ physical elements, like the steering wheel for a car, the bathroom faucets or the buttons for an elevator that we mentioned.

These traditional interfaces function primarily as mediators between the human and

the artificial system[1]. They translate intentions through clicks, text, or voice commands into inputs that the system behind the interface can process, and they transform the computations of that system into something we can understand and interact with.

The narrow definition is sufficient for addressing traditional design challenges. However, as artificial systems evolve in complexity and capability, our understanding and the evolution of interfaces must similarly grow along three dimensions, mirroring the development trajectory of artificial systems. Below, we look into these dimensions under which we classify the possible paths of evolution for the interfaces of artificial systems.

Functional expansion: Large language models (LLMs) underlying conversational agents (CAs) have evolved from relatively simple text prediction tasks, but nowadays reply to some of our queries with results that exhibit pattern recognition, planning and problem-solving at levels that were thought to require human-like cognition just a few years ago [2]. The academic community remains divided on how to interpret

1. Application Programming Interfaces (APIs) are another example of narrow interface. However, in this case they do not connect the user with the system but rather allow software systems to communicate with each other, like payment gateways processing transactions or mapping services calculating routes. Protocols are another of such examples of technological interface between artificial systems. For instance, while Anthropic's Model Context Protocol (MCP) to standardize how applications communicate and provide context to LLMs (proposed "like a USB-C port for AI applications" [4]) enables complex interactions, it still functions through a specific, defined channel (structured text with XML-like tags) and follows particular formatting conventions and rules. Even though the content and meaning transmitted through it can be more rich and varied than in the case of most protocols, it is relatively constrained to a narrow interface as far as **how information is exchanged**.

2. The trajectory from the IBM's classic chess-playing supercomputer Deep Blue (1985-1997) to DeepMind's successive game-players illustrates the progression of AI from domain-specific to increasingly general capabilities.

- AlphaGo (2016) required specific knowledge of the rules, the domain and human game data in order to master Go.
- AlphaGo Zero (2017) learned tabula rasa through self-play, only with the rules of the game as knowledge.
- AlphaZero (2018) generalized this approach to master three different games (Go, chess, and shogi) using "a single algorithm for all games" given only the rules [6].
- AlphaStar (2019) performed real-time strategy gaming (achieving Grandmaster level in StarCraft II), a domain that involves incomplete information processing, multi-scale decisions, and exploration and action

these capabilities [3]. Some scholars claim that the "reasoning" and other cognitive functions seemingly displayed by artificial systems are not really so, and are at best no more than an ill-conceived anthropomorphism of sophisticated statistical simulations, and at worst a manipulative branding strategy, rather than genuine cognitive functions. Others suggest the expansion of these capabilities represent emergent skills that warrant reconsideration of our definitions of reasoning (and cognition) itself. What remains apparent, regardless of our theoretical position, is that the functionally parallel range of these systems is significantly expanding, even if their underlying mechanisms differ substantially.

Modal expansion: CAs have undergone a remarkable transition from the "uni" to the multimodal and multilingual. Early systems were very limited to text-based interaction, preferably in English, without typos, and case-sensitive, and many of their lim-

---

selection in human time-scales [7].

- Finally, MuZero (2020) represented another step, learning the rules of the game from scratch through interaction, which allowed it to navigate environments with unknown dynamics such as Atari games [8].

This progression exemplifies the continual redefinition of intelligence itself, as each breakthrough shifts the **threshold for what constitutes intelligence**. From "AI will be intelligent when it can play chess" to "when it can play Go" to "when it can learn without rules," continuously expanding our functional definitions as systems achieve previously human-exclusive capabilities. As Demis Hassabis pointed this year "we do not have systems yet that could have invented general relativity when Einstein did with the information that he had available at the time", and "Can [AI] invent a game like go? Not just play a great move like move 37, or build AlphaGo that can beat the world champion, […] invent a game that is as beautiful, aesthetically, and so on, as go is?" [9] [10].

3. As the philosopher of mind Daniel Dennett might note, this creates an interesting "intentional stance" problem [11]. This refers to our tendency to interpret and predict behavior by attributing beliefs, desires, and rationality to an entity, essentially treating it as if it has intentions. So when we interact with sophisticated AI systems like large language models, we might instinctively slip into treating them as having genuine understanding, beliefs, and intentions. We might say an AI "thinks," "knows," or "wants" something, naturally attributing and **linguistically granting** cognitively analogue capacities to these artificial systems. Simultaneously, we question whether these systems can develop such cognitive capacities without subjective experiences or a sense of self, and an entirely inorganic base. Dennett argues that the intentional stance is primarily a predictive strategy: if treating something as having intentions helps us predict its behavior, then the stance is useful regardless of whether the entity actually has intentions in some deeper metaphysical sense. In any case, these systems are designed to mimic human-like responses, creating a situation where the intentional stance becomes increasingly smooth, even as philosophical questions about genuine understanding remain unresolved.

itations today still stem from that nature. This created a constrained communication channel similar to speaking with someone who can only read and write (reminding us of Searle's Chinese Room thought experiment [5]), but cannot have a glimpse into any other sort of digital file, much less generate its content. Gradually since around 2021, these systems have expanded to handle multiple modalities and formats including images, audio and video, and increasingly complex and diverse language. Contemporary systems can maintain spoken conversations, browse the web, process multiple file formats, analyze data visualizations and tables, collaboratively write or code, and even guide us in the physical world through a real-time camera feed, augmenting what we can do. For example, to find our glasses. This modal expansion mirrors aspects of human cognitive development, where language and multimodal processing enable us a broader range of interactions, and a richer conceptual model of the world. However, there is an inherent **asymmetry**. On one side, while vision and audition have reasonable digital analogues, other human senses like touch, smell, and proprioception (body movement and strength) remain challenging to digitize effectively. On the other side, and even more intriguingly, artificial systems might eventually develop capabilities that do not map onto human sensory systems at all, processing data streams that have no direct human perceptual equivalent, like a bat is echolocation. This creates what cognitive scientists following biologist Jakob von Uexküll might call different "**umwelts**" or experiential worlds between humans and artificial systems, with interfaces serving as the main translation layers, converging or diverging our perceptual realities.

Embodiment expansion: The progression from digital to physical interfaces probably represents the next profound shift in how we interact with artificial systems. Despite this level of expansion being still in its infancy, the underlying models behind our web browsers and phone apps are increasingly being integrated to steer physical devices ranging from semi-autonomous cars to humanoid robots. Making the artificial systems' interface physically present endows them with the ability to move through and manipulate the physical spaces we inhabit. The humanoid robots being developed by various research institutions and companies [4] signal this direction: from virtual to physical, and from disembodied text to embodied presence. This is only possible thanks to the modal expansion, as physical bodies of any kind require sensory capabilities. For example, a robot needs to perceive weight and space to ma-

4. Companies like *Boston Dynamics*, *Figure AI* and its Vision-Language-Action model, or *World Labs* and its aims to research spatial intelligence and Large World Models (LWMs).

nipulate objects effectively, which are traits that purely digital systems can mostly be ignorant about. The interface, in this context, expands beyond screens and speakers to include every physical aspect through which the artificial system interacts with the world and with us, like its appearance down to details like blinking, or the fluidity and pace of its movements. As these systems become more physically and perceptually sophisticated, the line between interface and embodiment blurs, raising deep questions about how we design for entities that share our physical reality.

What we call a **(2)** broader view incorporates (artificial) embodiment and invites us to consider physical embodiments themselves as interfaces. Just as our bodies serve as the interface between our minds and the physical world, translating intentions into actions, or sensory input into perception, robotic bodies can be understood as interfaces between artificial cognition and physical reality. This perspective draws from 4E cognition approaches, which suggest cognitive processes are fundamentally shaped by the physical form through which an agent experiences and acts upon the world.

In this sense, broad interfaces such as biological bodies and robots, generally involve:

- **Physical embodiment** across different modalities.
- Multiple **simultaneous** channels of interaction.
- **Integration** of diverse **sensory and motor** systems.
- Blurrier boundaries between the system and the interface.

Rather than viewing embodiment dichotomously (yes/no), we can consider this broad view on a spectrum. Even among humans, we find very significant variation in our embodied experience. While some people have highly developed proprioception and body awareness, some of us occasionally have a more forgetful or tenuous connection to our physical form (especially when in the flow of writing a master's thesis). Similarly, artificial systems can exist on a spectrum:

- Text-based interfaces represent minimal or absent embodiment, limited to linguistic expression and the embodied metaphors of language.
- Multimodal systems develop this embodiment to include visual and auditory processing.
- Robots and other physical interfaces flesh out a more developed embodiment by (inter)acting directly within the physical world.

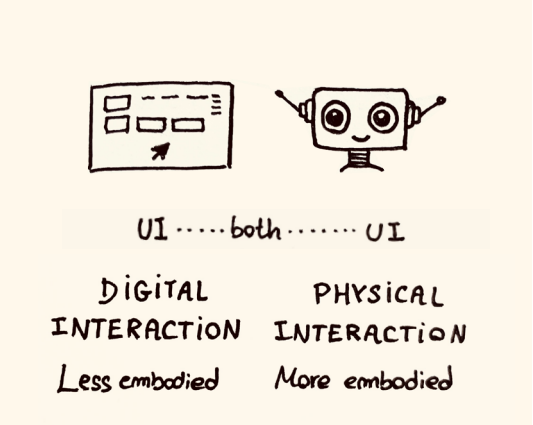


Figure 1. *From narrow to broad interfaces.* Both virtual graphic user interfaces and physical robots function as interfaces (meeting points) for human and artificial cognitive systems. Their degree of embodiment, instead of a dichotomy (yes/no) exists on a spectrum defined by multimodal capacities, simultaneous interaction channels, and the integration of sensory-motor systems. Original illustration by the author.

This spectrum view helps us understand interfaces not as fixed categories but as varying manifestations of the meeting point between cognition and environment. It also suggests that our design approaches must evolve as interfaces move along this spectrum, adapting to different degrees of embodiment and the specific challenges each will present. Developing such adaptable approaches requires a robust theoretical understanding for the analysis and design of communicative artifacts for cognitive systems (whether human, artificial or hybrid) across the full spectrum of embodiment possibilities.

⚑ Narrow UIs, Dimensions of Interface Expansion and Spectrum of Broad UIs

Narrow interfaces include GUIs (buttons, menus), VUIs (voice commands), CLIs (terminal commands), and tangible interfaces (physical appearance of objects).

Interface expansion occurs across three dimensions: **functional** (expanding cognitive capabilities from simple to complex problem-solving), **modal** (evolving from text-only to multimodal interactions including visual, auditory and multilingual capabilities), and **embodiment** (progressing from digital to physical manifestations like robots).

Broad interfaces include physical bodies and robots, and they involve physical embodiment across modalities, simultaneous interaction channels, sensory-motor integration, and blurred system-interface boundaries. This is not dichotomic yes/no, but rather they can exist on a spectrum.

2.2 Mediators and Constituents

Now that we have set the narrow and broad views of interfaces, and conceptualized the three dimensions of interface expansion, a more constitutional question emerges: what exactly is the relationship between interfaces and the cognitive systems they connect? Are they merely bridges between distinct cognitive systems, or do they become integral parts of cognition itself?

This question, drawing from cognitive science, leads us to distinguish between:

Interfaces as mediators: Interfaces function as channels between cognitive systems, translating between them while remaining separate from either.

Interfaces as constitutive: Interfaces may become what realizes the cognitive processes themselves.

To imagine this, we can think of bread. Flour, water, and yeast constitute the bread. Without them, there is no bread. The oven, however, mediates the transformation from dough to bread; it enables but is not part of what we consider "bread" itself. We could have, for example, bread on a stick over a campfire.

The distinction also leads us to consider: mediators or constitutive of cognition, but of whose? For humans, interfaces typically begin as explicit mediators but may gradually integrate into our cognitive patterns with prolonged use. A classic example most of us have seen is how smartphones have changed human memory strategies. How many phone numbers or map routes can you recall from memory compared to half your lifetime ago? Our cognitive patterns adapt to our tools, yet the tools themselves remain distinct from our cognitive processes. The interface has not become our memory, but it has certainly reshaped how we deploy our cognitive resources.

For artificial systems, interfaces may be more inherently constitutive, as these systems' "understanding" of the world is fundamentally shaped by their input and output channels. A text-only AI experiences a fundamentally different world than a multimodal system capable of processing images, audio, and text. The interface does not just connect the AI to the world, it largely defines what "world" exists for the AI, like viewing a landscape through a kaleidoscope that both limits and colors what can be perceived.

The etymology of "artificial" (from *artifex*, craftsman or master of an art, stemming from *ars* and *facere*, art + to make [12]) reminds us that artificial systems are meticulously designed and constructed, while human bodies became what they are through evolution and development. The awe-inspiring complexity of human biology, that we function at all, entails we have limited control over our own embodied interfaces. Medicine unveils these design limitations, while prosthetics and implants hint at more permeable boundaries between body and technology. For artificial systems, though, these boundaries are inherently more flexible and customizable, making the question of interfaces as mediators or constituents especially significant as we design them differently depending on how we answer it.

⚑ Mediating ≠ Constituting cognition

Both mediating and constituting elements are necessary for cognition, but while constituting elements realize the cognitive process themselves (such as flour, water and yeast for the bread), mediators only participate in the process (such as a modern oven).

This even page is
intentionally left blank.

3. *Why.* Of Noise and Experience in Communication

Are interfaces a communicative artifact? Is interface design a communicative act? What about interaction with CAs? To answer these questions, we need to challenge our assumptions about communication itself. How to break down what happens in a communication system? Where do interfaces sit in this system? Can we examine this question at multiple levels? How does the meaning of a message persist or transform as it crosses the boundary between all of those elements? Why does meaning so often get lost in translation in communication systems between the human and the artificial?

Communication models analyze the seemingly intuitive act of communicating, and provide a visual and conceptual representation of basic structures and processes describing how information flows between source and destination. By deconstructing this phenomenon into its constituent elements, we identify the components, barriers that influence whether understanding occurs, and concepts of relevance from a cognitive science lens.

These models represent a rare interdisciplinary confluence of engineering, design and philosophical inquiry converging to systematize the mechanics of meaning exchange. Like showing the roots before examining the tree, exploring communication theory reveals the invisible infrastructure underlying all communication, both in human-machine dialogue and interface design itself.

3.1 Shannon-Weaver's Model

Is there a thread that runs through the actions of talking with a friend, instructing a robot, typing to an AI, writing a letter or designing a presentation? Warren Weaver's early definition of **communication** as "all of the procedures by which one mind may affect another" [13] suggests they are all communicative acts.

The Shannon-Weaver model, developed by Claude Shannon and Warren Weaver in 1949, emerged from information theory and mathematical approaches to communication. Initially conceived by Shannon one year earlier as a framework to address technical problems in signal transmission [14], this model approached communication as a linear, mechanistic process focused on the accurate transfer of information.

Its origins in engineering and mathematics reflect the post-war scientific paradigm that sought to quantify and systematize communication.

Weaver translated the model to interdisciplinary audiences, extended the definition of communication to broadly include "procedures by means of which one mechanism (say automatic equipment to track an airplane and to compute its probable future positions) affects another mechanism (say a guided missile chasing this airplane)". Thus, we can summarize their idea of communication as "**procedures** by which one mind or mechanism affects another". He considers communication involves not only written and oral speech, but also music, pictorial arts, theatre, ballet, and "in fact all human behavior".

3.1.1 Core Concepts

The communication system considered in Shannon and Weaver's paper "The Mathematical Theory of Communication" (1949) may be **symbolically represented** as follows [13]:

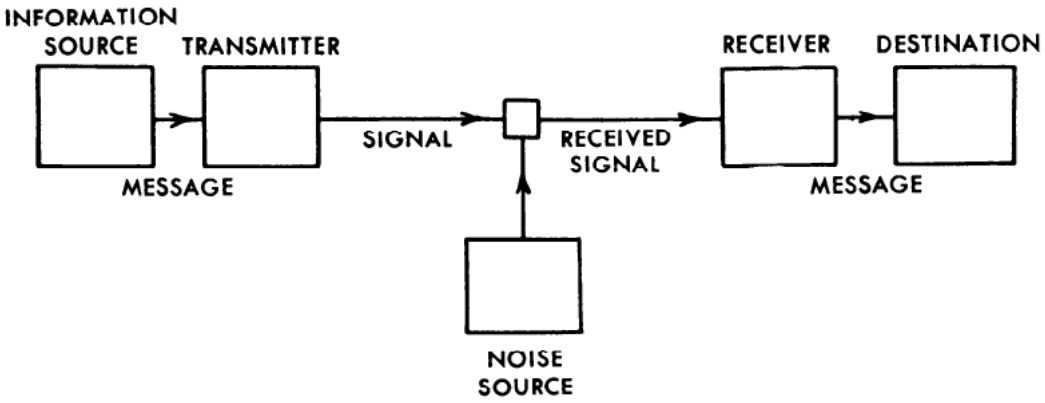


Figure 2. *Diagram of Shannon-Weaver's model of communication.* Linear transmission of a message from an information source through a transmitter, across a channel prone to noise, to a receiver and finally the destination. Developed to examine technical communication, these elements help to distinguish different components of communication between human and artificial systems. Original from Shannon and Weaver (1949).

This diagram illustrates the fundamental components and the flow of information in a general communication system. Let us define them, starting by the four square boxes on the top:

- An **information source** begins the process by producing a message, or sequence of messages, to be communicated to the receiving terminal. This message is selected out of the set of all possible messages[5], and as we have seen, can take very diverse forms, from written characters and spoken words to pictures or music.
- A **transmitter** "operates on the message in some way to produce a signal suitable for transmission over the channel". In general, this operation performed by the transmitter can be referred to as encoding[6]. What this transformation entails depends on the communication medium, and some of the examples below evince.
- A **receiver** does the opposite operation, using the signal received to reconstruct a message (decoding).
- A **destination** is commonly a person or thing for whom the message is intended. In Weaver's words, "when I talk to you, my brain is the information source, yours the destination; my vocal system is the transmitter, and your ear and the associated eighth nerve is the receiver" [13].

The **signal** can thus be understood as the encoded **message**, prior to being decoded. Or, in other words, the message transformed and constrained to a format in which it can be communicated (through a channel). A channel is the medium through which the signal travels from the transmitter to the receiver. Technical examples of channels include a band of radio frequencies, a pair of wires, a coaxial cable, or a beam of light. Examples in the paper to understand the relation of signal and channel include:

---

5. The concept of **information** receives special attention in their paper, and they emphasize is not the "normal usage of the word", and neither a synonym of "meaning". In fact, two messages, one meaningful and the other nonsense, can be equivalent from the perspective of information as defined in their theory. For example, the bits 0 and 1 might stand for two different messages, 0 being a full Shakespeare's play, and 1 being "yes". Thus, information is defined as a "logarithm of the number of choices" (for those more mathematically inclined), and as "a measure of one's freedom of choice when one selects a message".

6. As the most elemental transmitters are thought to only produce a signal, Shannon and Weaver introduce the concepts of encoding and decoding much later in their paper than their core elements, in reference to codes and cyphers, and noticing that when sophisticated transmitters and receivers possess **memories**, "the way they encode a certain symbol of the message depends not only upon this one symbol, but also upon previous symbols of the message and the way they have been encoded."

- In oral speech, taking the brain as the information source, the vocal system (transmitter) produces varying sound pressure (signal) that travels through the air (channel).
- In telephony, the telephone (transmitter) converts the varying sound pressure of the voice into a varying electrical current (signal) on a wire (channel).
- In radio, an electromagnetic wave (signal) is transmitted through space (channel).
- In telegraphy, the transmitter encodes written words into sequences of interrupted currents (signal) of varying lengths (representing and represented by dots, dashes, and spaces) to send them through a wire (channel).

Note that there is one more box in the diagram, essential and uniquely introduced by Shannon in this model. During transmission, the signal is often affected by unwanted additions or modifications collectively termed noise. The **noise source** symbolically represents these disturbances acting on the transmitted signal, resulting in a received signal that may differ from the original signal encoded by the transmitter. Noise can manifest in various forms, such as distortions of the voice in a telephone call, static in the radio, distortions in the shape or shading of a picture in television, or errors in transmission in telegraphy and typing.

The path of communication flows from the information source, through the transmitter, over the channel, affected by noise, to the receiver, and finally to the destination.

Now that we have the language, what does this entail for cognitive science, conversational agents and their interface design?

3.1.2 Weaver's Three Levels of Communication Problems

Weaver poses three questions, from which stem three corresponding levels and communication problems that will help us better frame the complexities of communication. We review them with our own examples:

- LEVEL A. How accurately can the symbols of communication be transmitted? (The **technical problem**.)
- LEVEL B. How precisely do the transmitted symbols convey the desired meaning? (The **semantic problem**.)
- LEVEL C. How effectively does the received meaning affect conduct in the desired way? (The **effectiveness problem**.)

The **technical problems** "are concerned with the accuracy of transference of sets of symbols (such as written speech), or of a continuously varying signal (telephonic or radio transmission of voice or music), or of a continuously varying two-dimensional pattern (television)" [13].

Imagine you are trying to whisper a message to a friend across a crowded room, or in a noisy bar. Level A would be concerned with whether your friend can even hear the individual sounds (the symbols) you are making, and if they are distinguishable from the environmental noise and not subject to signal loss. To understand this, we can also think of a phone call with poor connection. When going through a tunnel while in a car, you might hear noise, crackling, or parts of the words being dropped or distorted. The technical problem is about getting the sounds (symbols) across from your mouth to your friend's ear, or even from your phone to theirs, without these errors. If the connection is so bad that the signal is lost and your friend cannot even make out the individual words, then Level A has not been successfully addressed. The signal, the message and the effect are all lost in transmission.

Meanwhile, the **semantic problems** are concerned with "the identity, or satisfactorily close approximation, in the interpretation of meaning by the receiver, as compared with the intended meaning of the sender".

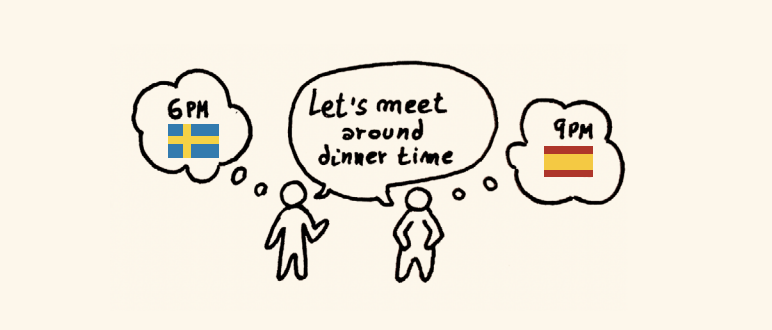


Figure 3. *Instance of a semantic communication problem.* This illustrates how perfect technical transmission (level A) of a message (to meet around dinner time) can fail if its meaning is interpreted differently by each communicator due to unacknowledged cultural conventions, resulting in misaligned expectations. Original illustration by the author.

Imagine you texted your friend from abroad "Let's meet around dinner time." In your culture, this means 6PM, but your friend comes from a culture where dinner is typically eaten at 9PM, as in my own. Though the message was technically transmitted perfectly (all symbols arrived intact), the meaning was misinterpreted due to cultural

differences. The semantic problem concerns whether the symbols you sent convey the meaning you intended, even when they are received perfectly[7].

The **effectiveness problems** are concerned with the "success with which the meaning conveyed to the receiver leads to the desired conduct on their part".

Imagine how a smoker perceives and comprehends the warning on a cigarette package. Their sensory systems accurately receive and decode the warning (no technical problem), and then semantically understand the meaning of "smoking causes cancer" (no semantic problem), or equivalent deterrents. However, if they are seeing the message it is likely that their behavior has not changed and they continue to smoke. This illustrates the effectiveness problem in which even if information is properly transmitted and its content understood, it does not have the desired effect, as other cognitive factors like temporal discounting (tendency to devalue temporally distant rewards [15]) or the neural reward pathways associated with addiction might prevent the sought behavioral response from occurring.

In short, the technical problem (A) is concerned with transmission, the semantic problem (B) is concerned with meaning and its interpretation, and the effectiveness problem (C) is concerned with the results or consequences of a message[8].

The levels we talked about in the previous section **grow on one another.** Level C, the effectiveness of a message, is only possible if B and A happen. One usually gets the desired actions or impact (the effectiveness problem) by getting the meaning

7. For an extended questioning and elaboration on meaning, semantic problems, the ambiguity and redundancy of language, Pfungst horses and Searle's Chinese Room, see our annex section "Semantic Problem of Pfungst horses and Searle's Chinese Room" at the link by the end of this document.

8. While some contemporaries criticized the effectivity level of Shannon and Weaver's theory claiming it equates communication with manipulation, they themselves state that we do usually communicate for a reason, whichever it is: *"it may seem at first glance undesirably narrow to imply that the purpose of all communication is to influence the conduct of the receiver. But with any reasonably broad definition of conduct, it is clear that communication either affects conduct or is without any discernible and probable effect at all. The problem of effectiveness involves aesthetic considerations in the case of the fine arts. In the case of speech, written or oral, it involves considerations which range all the way from the mere mechanics of style, through all the psychological and emotional aspects of propaganda theory, to those value judgments which are necessary to give useful meaning to the words 'success' and 'desired' in the opening sentence of this section on effectiveness."* [13] In other words, effectivity could merely be to evoke a smile in your interlocutor, or their enjoyment over reading your haiku.

across (the semantic problem), which in turn depends on actually getting the message and its signal across (the technical problem). However, the real world is messier than the theory suggests. For example, one might be able to reconstruct the meaning from a partially degraded message (such as a damp or half-burned letter) or a partial signal loss (the fragments of voice that reach you in the middle of a noisy bar, or a phone call with connection cuts). Language itself works partially as a web of pointers whose references get reconstructed in context [16] [17]. As Weaver notes, it is "highly suggestive for the problem at all levels that error and confusion arise and fidelity decreases" [13] but meaning and effectiveness can sometimes find a way despite technical degradation.

In the next section, we will address a limitation of this model (its treatment of meaning) which will lead into a necessary discussion of the nuanced distinction between signal and message. The next section will also include a discussion of the concept of noise and nested layers of communication. This connection will be important as we will be introducing the Schramm model of communication in sub-chapter 3.2.

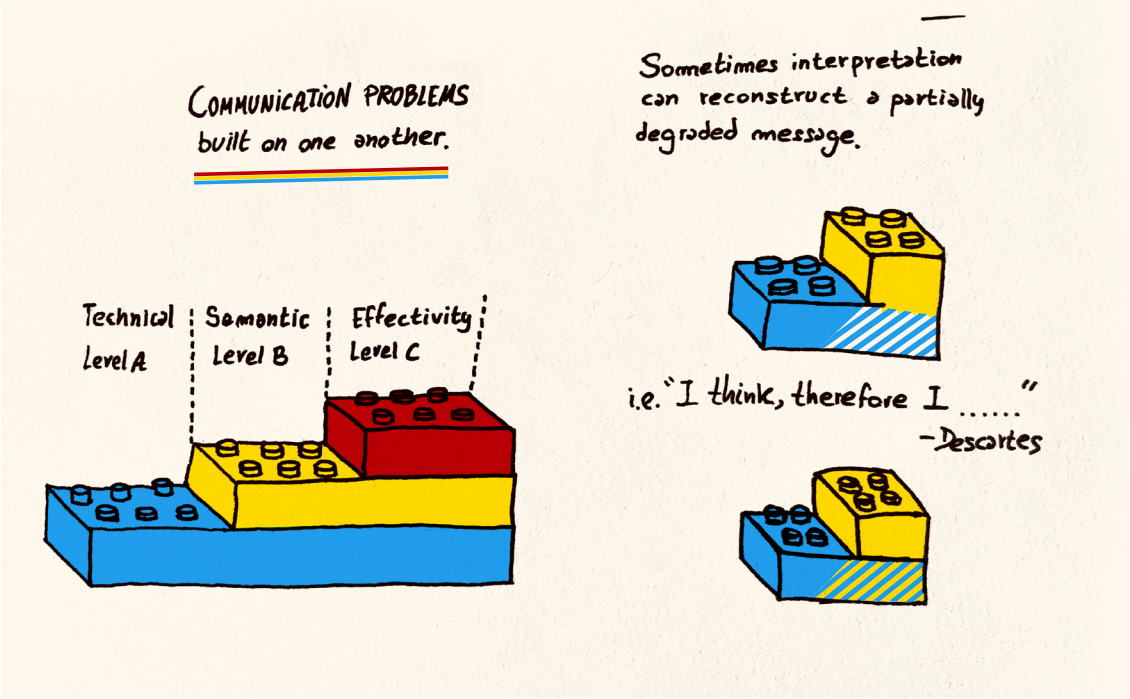


Figure 4. *Weaver's three levels of communication problems.* Left, visualized as interlocking blocks, the effectiveness of a message (Level C) is usually dependent on successful semantic interpretation (Level B) and technical signal transmission (Level A). However, as shown on the right, other levels can sometimes compensate, such as when interpretation and context may reconstruct meaning despite a partially degraded signal. Original illustration by the author.

3.1.3 Three Limitations of the Shannon-Weaver's Model

Now that we are familiar with the Shannon-Weaver model, its core concepts and the three levels of communication problems posed by Weaver as well as their inter-relatedness, we are finally at the point of exploring the limitations of the model. The first of the limitations relates to the limited treatment of meaning which leads to an ambiguity with regards to the concepts of **(1a)** signal and **(1b)** message. **(2)** The second limitation relates to the radical increase in complexity when we shift from human-human or machine-machine communication towards human-machine communication. For the latter, the concept of non-technical noise becomes a non-negligible aspect. **(3)** Finally, the last limitation we will touch upon in this sub-chapter relates to the linearity of the Shannon-Weaver model and the lack of treatment for nested layers and feedback involved in communication.

While Shannon-Weaver's model revolutionized information theory, it fundamentally conceptualizes communication as a linear signal transmission problem rather than a meaning-making process. Most of the second part of their paper, the core of Shannon's theory (which is beyond the scope of what will become relevant in this thesis) is concerned with the definition of information transfer and its mathematical formulation (Level A). Shannon explicitly states that "the semantic aspects of communication are irrelevant to the **engineering problem**" (Level B) [13]. Despite its genius, this mechanistic approach fails to capture most of the nuanced, contextual, and interpretive aspects concerned with interaction with human cognition and AI systems.

Although Shannon's initial mathematical theory explicitly excludes semantic considerations, Weaver recognized this and self-critically comments Level A is "a relatively superficial one, involving only the engineering details" while levels B and C, more concerned with the rest of this thesis, "seem to contain most if not all of the **philosophical content** of the general problem of communication" [13], which leads to fundamental misalignments that go deeper than just transmission "noise".

By focusing on information instead of meaning, Shannon's development of information theory willingly neglects the fundamental **ambiguity and relativity of language** (what do words stand for?) and the relevance of the **environment**, the **interlocutors and their experiences** in a dialectic communication. We will follow up on this in the following section, when we introduce Schramm's model of communication which brings up and partially conceptualizes all of these tricky nuances.

Because of the limited treatment of meaning just commented, and while they are more concerned about information, one thing Shannon and Weaver do not put sufficient emphasis on, but which we particularly find helpful is the distinction between signals (the technical carrier) and messages (the meaning triggers, so to speak), and the clarification that **Signal does not equal Message**.

(1a) Now, what is a signal? A signal occurs just between the transmitter and the receiver (Figure 5 top), and it is rather a technical thing, a carrier. It is how a message is sent. Be it the "varying sound pressure" of a voice in front of us, the "varying electrical current" of its conversion to audio for a phone call, the electromagnetic wave of a radio, or the vibrated tappings of Morse code. For semantic purposes, the transmitter is what encodes the message into a signal, and the receiver is what decodes the signal into message.

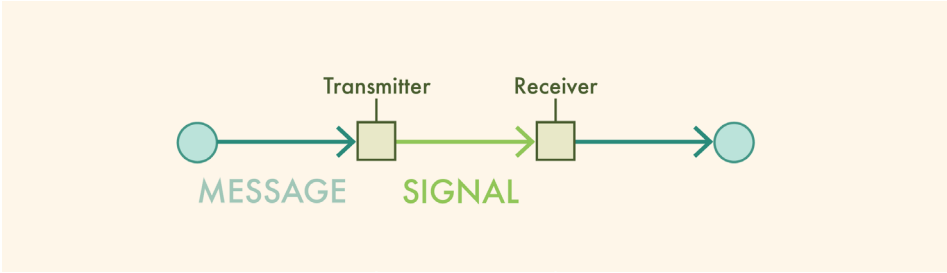
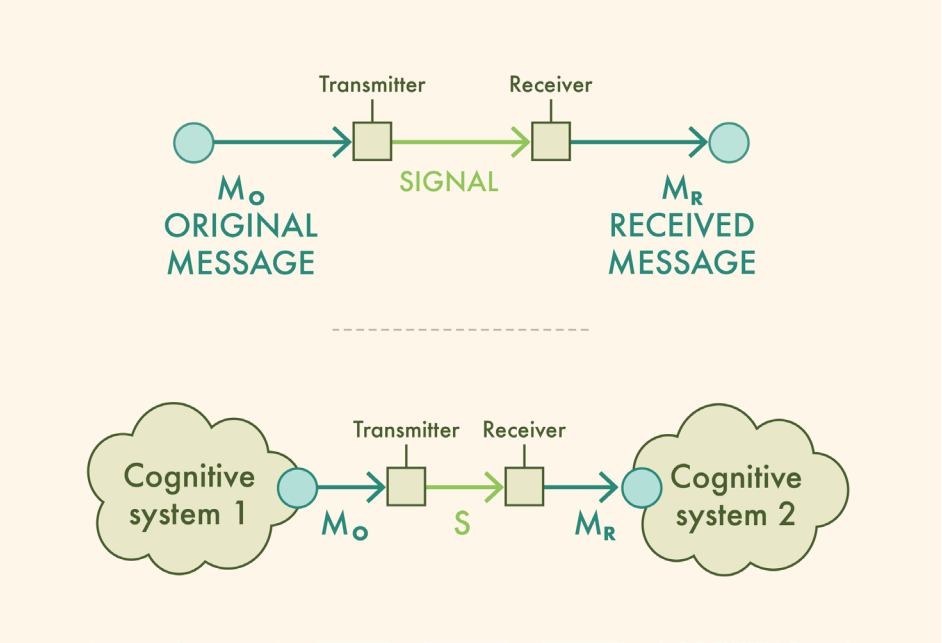


Figure 5. *Illustration of the limited domain of the signal* (top), which only captures what travels between transmitter and receiver. The expanded view (bottom) reveals the preceding step of message formation and the subsequent step of message reconstruction, both situated within the cognitive systems of sender and receiver. Original illustration by the author.

(1b) What is a message? A message, in turn, is what this signal was before being encoded (by the transmitter) and after being decoded (by the receiver) for the destination (Figure 5 bottom). Although Shannon and Weaver diagram only distinguishes between signal (pre-noise intervention) and received signal (post-noise intervention), we would like to add that it could be helpful to distinguish too between the message (intended, original, between the source and the transmitter) and the received message (between receiver and destination).



In Figure 5 above, we see the original message (M_O) transformed into signal (S) and received message (M_R). This signal-message distinction suggests the multi-layered nature of communication in digital interfaces. If we imagine an interaction with a CA and its interface, from a user's perspective we can distinguish between:

- Message formation: From the user's intention and their cognitive system to generating a consistent original message.
- Signal transmission: When the user encodes the message into physical actions (typing, gestures, voice commands etc.).
- Signal reception: The cognitive system at the other end, in this case the CA, receives these inputs as electrical signal.
- Message reconstruction and interpretation: The cognitive system at the other end, in this case the CA, must decode these signals to reconstruct the intended message.

These distinctions help us understand that both dialectic communication and interface design for CA are not merely about signal transmission but also about facilitating accurate message reconstruction across this cognitive-physical-digital boundary, and back. In a way, we can say that it is as much about the signal as it is about the signs[9], a concept to which we will come back in Schramm's model of communication

9. **Signs** are something semiotics studies theoretically, while design applies into practice. A sign is a meaningful unit that represents or stands for something other than itself (it can be a color, a shape, a word) functioning as the basic element in systems of communication and interpretation. **Semiotics** is the systematic study of signs

in sub-chapter 3.2.

(2) The second limitation of the Shannon-Weaver Model can be formulated as follows: As the model was developed during the post-war period, it stems from an interest in machine-machine communication and with an explicit focus on signal transmission which eliminates many layers of complexity involved in human-computer interaction. Unlike human-to-human communication, or machine-to-machine communication, where both parties share relatively similar cognitive backgrounds, the "noise" in human-AI interactions stems from radically different systems trying to communicate. Now, as the careful reader might have recognized, the concept of noise has not yet been sufficiently dealt with. The underlying question for our purposes is: Where is the Noise?

In the mathematical theory, noise refers only to unwanted additions or distortions to the signal during transmission. While this is relevant to the technical underpinnings of interfaces and the dialectical nature of CAs themselves, the concept does not fully capture the "noise" of ambiguity such as poor usability or confusing language that can impair usability at the semantic level, and user experience at the effectiveness one. For us, **noise** will be an umbrella term for anything that may distort the communication flow, altering the signal or the traceable message, and prevent it from being interpreted or reach its destination effectively at any stage (which has relevant implications for interpretability, cognitive science and design, as we will see throughout the thesis).

Weaver assures the reader that "it is almost certainly true that a consideration of communication on levels B and C will require additions to the schematic diagram" (level B relating to the semantic one and C the effectiveness), and theorizes one can imagine adding boxes for a semantic receiver and a semantic noise to the symbolic representation (diagram) of communication:

---

and sign processes, examining how meaning is created, communicated, and interpreted across various contexts and media. In a way, semiotics provides a philosophical basis explaining why visual elements can communicate specific meanings, enabling designers to purposefully manipulate form, color, typography, and other elements in an interface, such as composition, to evoke intended interpretations and emotional responses from users or viewers. Arguably, all language is made of signs too. As this is a Pandora's box in itself, and deserves more space than a footnote in this thesis allows, for now we just refer the reader to linguists such as Ferdinand de Saussure, and philosophers such as John Locke and Charles Sanders Peirce.

> *One can imagine, as an addition to the diagram, another box labeled "Semantic Receiver" interposed between the engineering receiver (which changes signals to messages) and the destination. This semantic receiver subjects the message to a second decoding, the demand on this one being that it must match the statistical semantic characteristics of the message to the statistical semantic capacities of the totality of receivers, or of that subset of receivers which constitute the audience one wishes to affect. Similarly one can imagine another box in the diagram which, inserted between the information source and the transmitter, would be labeled "semantic noise," the box previously labeled as simply "noise" now being labeled "engineering noise" From this source is imposed into the signal the perturbations or distortions of meaning which are not intended by the source but which inescapably affect the destination. And the problem of semantic decoding must take this semantic noise into account. It is also possible to think of an adjustment of original message so that the sum of message meaning plus semantic noise is equal to the desired total message meaning at the destination.*

Although Weaver suggests the idea of "semantic noise" as a perturbation of meaning and "semantic receiver", those are not part of the original mathematical treatment of noise, nor extended or included in the symbolic representation of the model. If we are interpreting Weaver right, we hypothesize the representation with a semantic receiver between an engineering receiver, and destination and semantic noise between the information source and the transmitter, could look like the following:

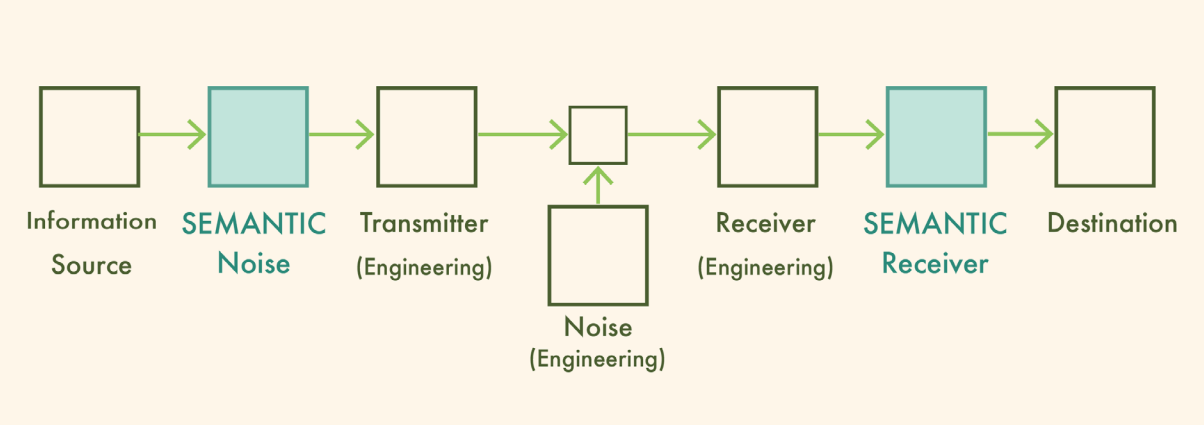


Figure 6. *Diagram extended as per Weaver's further research suggestions*, not diagrammed in the original paper. This representation hypothesizes the placement of semantic noise (between source and transmitter) and a semantic receiver (between engineering receiver and destination) within the original model. Own representation based on Weaver's suggestions.

This symbolic representation falls short both in design and multiple conceptual aspects we are discussing and elaborating on:

- How to find similar but distinguishable symbolic ways to draw signal and message?
- How to distinctly represent the level B on which semantic problems take place?
- How to add the processes of encoding and decoding without cluttering the diagram?
- How to imply semantic problems are not restricted to the receiver, but extend to the sender too?

We see eye to eye with Weaver in that there is a semantic receiver and semantic noise that must be taken into account, especially when dealing with human-machine interaction instead of human-human or machine-machine interaction. However, how to evince different moments, ways and levels in which noise might arise and distort or interrupt not the signal, the message, and the very same processes (encoding, decoding, transmitting, receiving)?

All components and processes, including the transmitter and the receiver, or external factors, may introduce more noise, and **information loss** or incongruencies. Provisionally, we will call these different kinds of noise:

- **Technical noise**: Literal distortions in data transmission, such as typos or connection issues.

- **Semantic noise**: Misinterpretations that occur because humans and AI systems have different experiential worlds, especially if a receiver is not equipped to decode a language (be it visual or text-based!). For example, while a CA might master the use of language, it might miss cultural or personal connotations it has for a user.

- **Effective noise and barriers:** When the message is received and understood, but constraints and distortions still apply, making it err on the task. A simple version of this is when a user specifies a task towards the wrong target. Effective noise is not necessarily always undesirable, as in the case of users asking a CA for something clearly prejudiced or violent, such as fabricating a bomb. In those cases, similarly to the smoker reading the warning on a cigarette pack but opting for not doing what the message intended to trigger, the CA understands it, but its guardrails derail the answer, in this instance for ethical reasons.

- **Conceptual model noise**: Perhaps the most consequential layer are these discrepancies (and unknowns) between how users imagine the AI works and how it actually functions. While this does not necessarily correspond to effective level, and it melts with semantic noise, it is worthwhile to distinguish, and it is possibly the most relevant component in most unsuccessful interactions between humans and CAs today. For example:

    - **Explainable AI (XAI)** efforts can be understood as attempts to reduce conceptual model noise by making the AI's processes more transparent and comprehensible to humans. We will continue to elaborate on this inside of the discussion of Norman's work applied to CAs, as a future direction dealing with the connection between XAI, interpretability and interaction design.

    - A **properly designed interface** can reduce conceptual model noise by aligning the user's conceptual mental model more closely with the system's actual functioning, as we will see in chapter 4, concerned with interaction design. Each interface design choice either amplifies or reduces this noise. For example, an interface that anthropomorphizes the AI might increase conceptual model noise by suggesting capabilities the system does not have.

Consider the children's game of the broken telephone, where a whispered message travels through a chain of people, increasingly distorted at every step. This playful scenario exemplifies how noise operates across all our identified levels simultaneously, technical distortions in the acoustic transmission, semantic misinterpretations as unfamiliar words get replaced with similar-sounding ones, and effectiveness failures when the final message bears little resemblance to the original intention. The game is a case in point about how errors accumulate and amplify through the communication chain: a single technical mishearing triggers further misinterpretations, creating a castle of cards where each error gives rise to the next. In human-AI communication, this multiplicative effect of noise across multiple layers of communication becomes even more pronounced, as the differences between human and artificial cognitive systems create additional opportunities for these errors to occur and propagate throughout the interaction.

This is why the element of noise, and its potential occurrence across the three levels,

is awfully important to keep in mind. In a visual format, a signal is represented as a straight-line arrow (standing for an uncommonly non-distorted signal) towards the receiver. As the addition or disruption of technical or engineering noise would alter the nature of that signal, we can imagine the original signal representation turning it into a zig-zag-line arrow (standing for a distorted signal). Semantic noise, however, deals with a different layer, and thus we represent this noise and the message with a different color, a spiralling line around a signal towards a circle that overlaps with the green box, to symbolize interpretation. Finally, effective noise, or effective problems, would rather be represented by the arrow erring direction, or being steered away. For now, we do not draw "conceptual model noise" yet, as conceptual models is something we will more thoroughly explore and represent in the section of Don Norman's fundamental concepts of interaction.

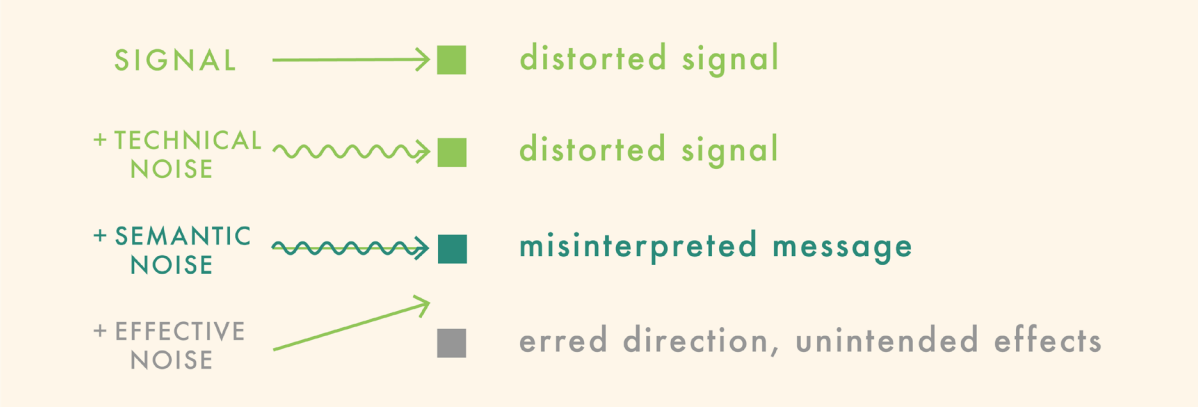


Figure 7. Visual representation of types of noise in human-CA communication, with distinct representations for each. Technical noise originates in Shannon-Weaver's model, semantic noise is suggested by Weaver as further research, and effectiveness noise is proposed here. A fourth type, relative to the conceptual model, is introduced in the text but will be further developed through Norman's design principles in Chapter 4. Original illustration and visual language by the author.

(3) One of the first things described in Shannon-Weaver's model of communication is how the information source begins the process by producing a message. However, we ask: Is it not a bit arbitrary where the end and beginning of communicative interaction lie? Where does communication start? The world is not a system with clear cut boundaries where one can, at given time, perfectly delineate a beginning and an end. We want to highlight that in a non-modelic scenario, especially concerning the digital world, there is a layered nature of communication that makes any interaction considerably more complex to represent.

Remember from sub-chapter 3.1.1 about core concepts, both oral speech and a telephone call are examples of communication. We can even imagine them as the same communication act, seen from different perspectives. Each has the key elements of the Shannon-Weaver's model including information source (produces the message), transmitter (encodes the message into signals), channel (the medium carrying the signals), receiver (decodes the signals back into a message), destination (where or to whom the message arrives) and noise (interference that can distort the signal). To illustrate this, let us identify in the examples those core concepts:

- For example, taking oral speech: The nervous system and thoughts of the speaker as the information source, the vocal apparatus as the transmitter (encoding neural signals into sound waves), air as the channel, environmental sounds as noise, and the listener's ear as a receiver, while their nervous system and thoughts are presumably the destination.
- Taking a phone call: The speaker as information source, the phone's microphone and encoding system as transmitter, the telephone network as channel, network interference or static as noise, the receiving phone's speaker as receiver, and the listener as destination.

In this case we can look at communication at the level of human-human interaction, or at the level of human-technology interaction. Where we set the information source, and how general or specific we are in it determines the analysis of the elements that follow. Taking the two examples together we have a case in point of nested communication: oral speech over a telephone call. In this case, there are several layers of transmitters and receivers, and the original message gets transmitted through different signals before reaching the person on the other side. As we see above, the information source is relatively displaced from the nervous system to the whole person depending on how we frame communication.

In these lines, the granularity of our analysis determines how many layers we identify and from which perspectives we examine them. How granular do we want to be? For instance, neuroscience reveals that speech production itself involves complex communication between bodily systems (respiratory, phonatory, and articulatory) with voice actually being the result from interactions between the central nervous system and vocal apparatus (including mouth, jaw, tongue, and vocal cords). However, from another perspective, in a telephone call, the voice constitutes the message prior to encoding, while the encoded voice becomes the audio signal transmitted through the channel. This multi-layered view is consistent with the model Shannon-Weaver, as

their model can be applied recursively at different levels of analysis and from multiple perspectives. And it entails that each "destination" in one communication process can become the "source" in the next one, creating a **chain of nested layers of communication**.

We have identified three core limitations of the Shannon-Weaver model. The next section serves as a synthesis of everything we have touched upon in chapter 3.1 and then we move on to discuss Schramm's model of communication in chapter 3.2.

3.1.4 Synthesis

To start this synthesis, we will provide some symbolic representations (diagrams) of **(1)** the different kinds of noise that we described in the previous section for the case of human-AI interaction. And then, **(2)** provide a final recap of chapter 3.1.

(1) To graphically represent similar but distinguishable signs for signal and message, as well as represent different kinds of noise at several stages, we use differences in the arrow lines, as sketched in the concept for the sections "Signal and Message" and "Where is the noise?". To distinctly represent the level B on which semantic problems take place we use a distinct color, namely turquoise. To add the processes of encoding and decoding without cluttering the diagram, and imply that semantic problems are not restricted to the receiver, we have included a new shape (circles in Figure 8) to the diagram, overlapping two circles as backgrounds of the source and the destination, but each in between the original squares and their corresponding messages, to more accurately reflect the principal role of these components for the semantic level. Finally, to represent a chain of nested communication, or multiple layers (e.g. of transmitters and receivers), we have just duplicated their representation overtime, overlaid upon one another respective to the message transmission, and alternatively colored to improve visibility.

Remember our discussion of the nested layers of communication. The most relevant information source-destination pairs for human-CA interaction to which we can apply the diagram in Figure 8 are (i) AI practitioners and designers -> users, (ii) conversational agent -> user and the (iii) the interface itself -> user. In the Shannon-Weaver model that inspired the above diagram, the information source is uni-directional. The next model we will explore in the next section, the Schramm's communication model, will be bidirectonal where the information flows from both sides. Before moving

on to the next sub-chapter, let us offer a brief recap of this sub-chapter.

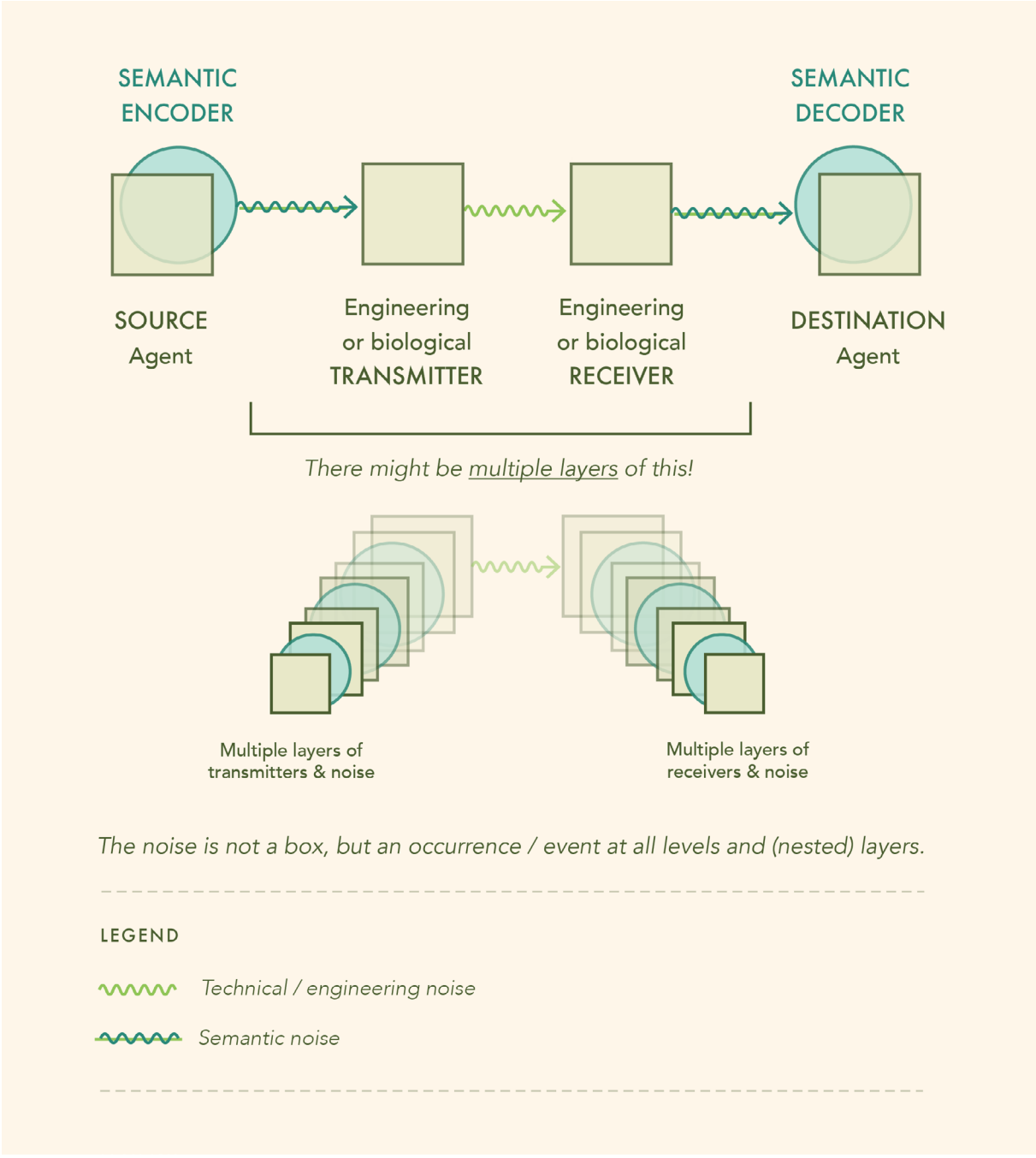


Figure 8. *Synthesis diagram* integrating extensions discussed throughout sub-chapter 3.1: semantic encoding and decoding as distinct from their engineering counterparts, multiple nested layers of communication, and noise as an occurrence at all levels rather than a separate component. Original illustration by the author.

(2) The Shannon-Weaver model, born from the crossroads of wartime engineering and Shannon's mathematical theory of communication, offers a linear, mechanistic

view of communication focused on the accurate transfer of information as a signal from a source to a destination, from a transmitter to a receiver, through a channel, potentially affected by noise. Weaver extended this by defining **communication** as procedures by which one mind or mechanism affects another, framing Shannon's theory from a wider lens, and making it more accessible. Furthermore, Weaver proposed **three levels of communication problems**: technical (A, accuracy of transmission), semantic (B, precision of meaning), and effectiveness (C, desired impact). We highlight the distinction between the **signal** and the **message**, emphasizing that the signal is merely the carrier.

While it sets the bedrock for understanding and structuring the basic components of communication, we address a significant limitation in the Shannon-Weaver model's **treatment of meaning**, as it primarily focuses the engineering problem of signal transmission (Level A), and largely neglects the complexities of human and AI interaction involving interpretation and context (Levels B and C). We also explore the concept of **noise**, which we do not limit to technical distortions like in the original paper, but extend it to semantic (misinterpretations), effective (failure to achieve the desired outcome) and conceptual model (failure to understand the other agent) **interferences**. The model also does not fully account for the ambiguity of language, the influence of environment and interlocutors, the nested layers of communication or the multiple kinds of noise. Attempts to evolve the model and its symbolic representation to incorporate semantic aspects, such as Weaver's suggestion of a "semantic receiver" and "semantic noise," indicate a recognition of these limitations. Although it remains unexplored in the original piece, we have attempted to explore and diagram some of these through the discussion.

Could there be models that go deeper into these very aspects, exploring how meaning is truly constructed and shared between communicators, especially in the intricate dance of human interaction? The stage is set for the next steps in our excursion into communication theory, and Schramm's model of communication will introduce us to perspectives that directly address these very nuances.

⚑ Signal ≠ Message, and noise is not only technical

In communication theory, **communication:** procedure by which one mind or mechanism affects another.

- **Message:** content selected by an information source to be communicated.
- **Signal:** encoded form of a message, transformed and constrained to a format suitable for transmission through a channel.

Three **levels of communication problems**:

- **Level A**, the technical problem (how accurately can symbols be transmitted?).
- **Level B**, the semantic problem (how precisely do transmitted symbols convey the desired meaning?).
- **Level C**, the effectiveness problem (how effectively does the received meaning affect outcome).

Noise can distort communication at any of those levels

→ Technical noise corrupts or disrupts the signal, semantic noise misrepresents intended meaning, effectiveness noise deflects the intended outcome, conceptual model noise (the distortion that accumulates when the receiver's understanding of the system itself is misaligned, proposed by us) impairs understanding.

Nested layers of communication: from a different perspective or granularity of analysis, any element of a communication process can play a different role (i.e. the human nervous system = destination for sound waves arriving at the ear, source for words we formulate, encoder and decoder of neural signals).

3.2 Schramm's Model

Wilbur Schramm's model in the paper "Procedures and Effects of Mass Communication" (1954) [18] represents an evolution in theoretical understanding, emphasizing the interpretive aspects of communication that Weaver hints at. Schramm situated communication within social contexts, and explores how meaning emerges through interpretation rather than mere transmission. This represents a significant departure from the more mechanistic Shannon-Weaver approach of the previous sub-chapter. Schramm's model addresses some of the limitations of previous models by introducing critical concepts such as **(1)** overlapping fields of experience and **(2)** feedback between communicating entities, an insight particularly valuable for human-AI interaction and cognitive science.

By positioning communication success as dependent on shared cognitive contexts rather than merely signal fidelity, Schramm provides a more accurate theoretical basis for understanding human communication, and why humans and AI systems frequently misinterpret each other despite perfect technical transmission of signals. With his model, we reconceptualize CAs' interfaces not simply as information conduits but also as messages, signs, and spaces where distinct cognitive systems with fundamentally different experiential worlds meet.

For Schramm, the Latin etymology of communication gives away its definition (*communis* is also present in words such as *community* or *commune*), and brings him to affirm that "when we communicate, we are trying to establish a 'commonness' with someone. That is, we are trying to share information, an idea, or an attitude" [18]. The roadmap for this sub-chapter, as was with the previous one, is to first introduce the core concepts (3.2.1), then to discuss what we consider the most important contributions of the model (3.2.2) and then discussing the limitations of the model (3.2.3) before closing with a synthesis (3.2.4). What will be different in this sub-chapter is that we will introduce related ideas from different authors that we believe contribute to our understanding of the Schramm model and illustrate how communication theory connects with cognitive science. We will do this in sub-sections 3.2.2.1 and 3.2.2.2.

3.2.1 Core Concepts

The communication system considered in Schramm's paper may be first **symbolically**

represented as follows [18]:

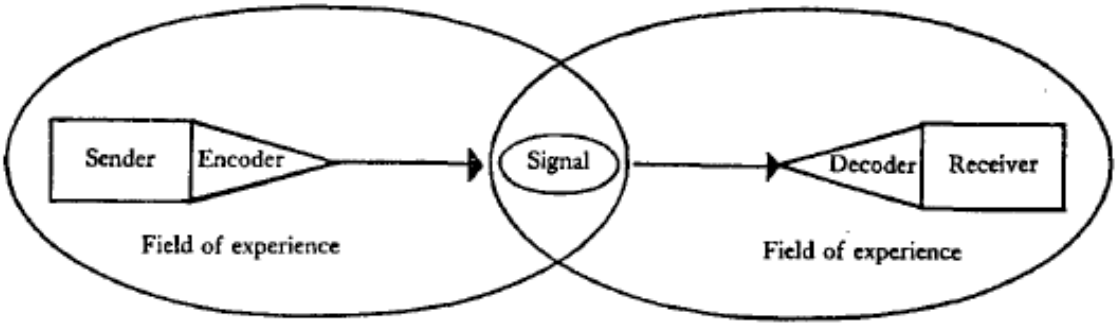


Figure 9. *Schramm's model of communication.* Encoding and decoding are situated within each communicator's field of experience, making them processes shaped by accumulated beliefs and culture. Original from Schramm (1954).

Schramm gradually increments the complexity of the model, to follow the flow of his process:

1. Start with a basic linear process (Sender -> Encoder -> Signal -> Decoder -> Receiver).
2. Then, integrate **encoder with sender** as a pack, and **decoder with receiver** as another pack.
3. Then, highlight the relevance of each's **fields of experience** (the two big circles and the phenomenon of overlap).
4. Take into account the factor of time.
5. Consider a message gets a response, and introduce **feedback.**
6. Examine more in depth what is within each communicator (the packs of sender > encoder and decoder > receiver), each of us, as a **decoder - interpreter - encoder** ball.
7. Consider a dialogue.
8. Expand it to societal and mass communication.

Many of the elements in this first representation, and their naming, draw from media theory of the times, such as Shannon-Weaver's model (sender, as a source, message, and receiver as a destination), but Schramm's frame of mind is more human-centric. A **sender** is the originator of the communication, "an individual (speaking, writing, drawing, gesturing) or a communication organization (like a newspaper, publishing house, television station, or motion-picture studio)" trying to "establish common-

ness" with a **receiver**, the individual or audience who hears, watches, reads or experiences the message. This **message** is basically anything the sender wants to share, put into a form that can be transmitted such as "ink on paper, sound waves, electrical impulses, gestures, or flags" [18].

This transformation process involves the steps **encoding** -> transmission -> **decoding**. While Shannon's focus was on the transmission (and its noise, Level A), Schramm's focus is on what happens at its ends (Level B). How do we transform mental states (what he calls "pictures in our heads") into transmissible signals? As he succinctly put it, these mental images "cannot be transmitted until they are coded" into language or other symbolic systems that can travel efficiently between minds (Level A) and effectively influence others (Level C). While for electronic communication he illustrates the encoder would be a microphone, and the decoder an earphone, at what we will call a different layer of communication (3.1.2, see the discussion of Shannon-Weaver's model and the proposal of nested layers of communication in the three levels of communication problems) Schramm considers the sender and encoder one person, while decoder and receiver another, and their signal to be language [18]. In this way, he elegantly **integrates sender and the process of encoding** and decoder and the process of receiving, without eliminating the distinction.

Schramm recognized the profound tradeoffs inherent in choosing different communication **media**. Spoken words remain limited in reach without technological amplification (they "can't travel very far unless radio carries them"). Written language sacrifices immediacy for permanence and range ("go more slowly than spoken words, but they go farther and last longer"), transcending even their creators "the Iliad, for instance; the 'Gettysburg Address'; Chartres cathedral", or his very same work. Though Schramm primarily discusses linguistic messages in these examples, his family background in music (his father played the violin besides being an attorney, his mother played the piano, and he played the flute [19]) and political science suggests his framework extends to other forms of **artistic expression** and political discourse as equally valid forms of coded meaning.

3.2.2 Schramm's Fields of Experience, Feedback and Interpretation

3.2.2.1 Experience and Environment: From Fields to Umwelts

As mentioned above, our aim in what follows is to **(1)** explain Schramm's concept of fields of experience and then **(2)** Uexkull's Umwelt.

(1) Schramm establishes that **messages** function through **signs** (like words), which are a **signal that references something in experience**, and whose meaning is learned through association. He illustrates this through the example of "**dog**", a linguistic sign that evokes varied mental representations depending on one's particular encounters with dogs. Critically, Schramm observes that signs (and thus, language) operate as experiential shorthand, offering portability (at the price of only partially capturing what they refer to), but no two individuals possess identical interpretive backgrounds for decoding these signs, as their experiential references inevitably differ.

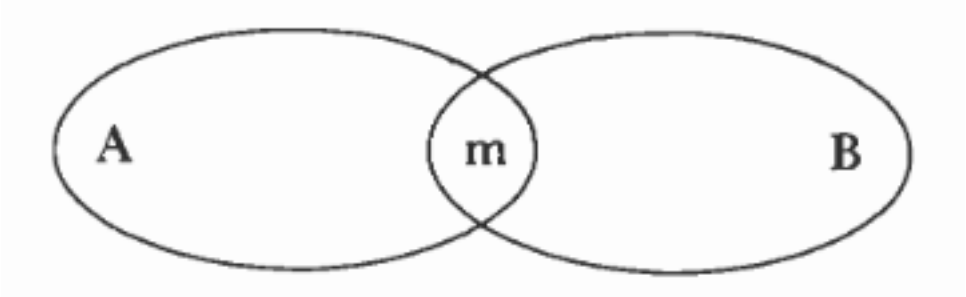


Figure 10. *Schramm's Overlapping Fields of Experience.* This diagram illustrates that communication is only possible in the shared area (m) where the field of experience of the source (A) overlaps with the field of experience of the receiver (B). Original from Schramm (1954).

His concept of **fields of experience** introduces a new dimension to communication theory by framing the exchange between sender and receiver as necessarily shaped by their respective accumulated experiences (being "in tune") [18]. Visualized as overlapping circles surrounding each communicator (aggregated sender-encoder and decoder-receiver), each of these fields in a way encompass everything that a person has lived through, learned, and internalized (knowledge, beliefs, values, and cultures).

According to Schramm, the larger the area where these circles overlap, the greater

the "commonness" or shared experience, and thus, the easier and more effective communication becomes. For instance, if two people have a similar cultural background or education, their fields of experience will likely overlap significantly, making it simpler for them to communicate with one another with relative ease. Conversely, if there is little or no overlap, communication becomes difficult or even impossible. Schramm illustrates this principle with the imaginable example of someone seeing an airplane for the first time, if they had never seen or heard of an airplane, they would only be able to decode the sight of it in terms of whatever experience they previously had, such as the sight of a **bird** [18].

This concept is consequential because without shared context or reference points, perfectly transmitted messages remain useless. The similarity of meaning that two communicators will perceive in a message depends on finding an area where their experience is sufficiently similar that they can share the same signs efficiently, which he calls "their fund of **usable experience**" [20].

We have seen that for Schramm, the field of experience is everything that a living being has experienced. Now, if we are to talk about the totality of experiences for an agent, we have to make it clear what this agent is actually capable experiencing, i.e. what is the environment they are embedded in and which aspects of the environment they are equipped to experience. This will be important because, depending on the structure and history of agents, their fields of experiences will differ. In other words, perceptions of one class of agents are different from those of another class of agents. For example, if one sends a colorful message to a destination that cannot perceive color, there will be no effective communication taking place.

(2) It is precisely for this reason that at this intersection, we believe that it is worthwhile to bring in Jakob Johann von Uexküll and briefly discuss his concept of Umwelt that will serve as a bridge between communication theory and cognitive science.

The important realization about the differing fields of experience is elegantly captured with the notion of "Umwelt" by Jakob Johann von Uexküll, zoologist and predecessor to biocyberneticism. He proposed already in 1909 in his work "Umwelt und Innenwelt der Tiere" - "Environment and Inner World of Animals" that each species inhabits not just a physical environment but a distinct perceptual sphere or **"surrounding-world" (Umwelt)** shaped by the organism's unique sensory and cognitive apparatus [21]. Because for Uexküll each creature perceives and interacts with its environment in a way dictated by its particular sensory organs and biological

needs, its accumulated experiences and understanding of the world (its Umwelt) will be distinct from those of other species. In our own words, that an oyster, a bat, and a human each live in fundamentally different experiential worlds despite occupying a common physical space. This is what Schramm illustrates with the overlapping circles and his image of a person seeing an airplane for the first time and being able to decode it only based on their existing field of experience, perhaps interpreting it as a large and noisy bird.

These umwelts (Umwelten) can be very different between different species or forms of life, leading to the respective "fields of experience" having minimal to no **overlap** with each other. For example, while bats "see" by sending out sound waves and mapping their world through echoes bouncing back (echolocation or sonar)[10], oysters, fixed in space, experience their environment mainly through feeling water pressure, movements and detecting chemicals. Humans, however, usually[11] rely heavily on very different senses, such as vision and hearing, within a nervous system that

10. The notion that also appears in Thomas Nagel's "What is it like to be a bat?" (1974) [22], which underscores the inherently subjective nature of consciousness and experience ("the subjective character of experience is inherently tied to a particular point of view"). Nagel argues that (assuming both the bats and ourselves have experiences), due to the fundamentally different sensory apparatus and way of interacting with the world (echolocation, for instance), we can never truly grasp the qualitative nature of a bat's experience, the **what it is like-ness**, and viceversa. Bats perceive their environment primarily through detecting the reflections of echoes bouncing back from objects within range, echoes of their own fast, modulated, high-frequency shrieks. Their brains are wired to correlate the outgoing impulses with the subsequent echoes, and it enables bats to discriminate qualities such as size, shape, motion or distance. and texture comparable to those we make by vision. Nagel also suggests that when we contemplate the bat's experience, we are in a similar position that intelligent bats or Martians would be in trying to imagine what it is like to be us ("our own experience provides the basic material for our imagination, whose range is therefore limited"). However, he also notes that the fact that we lack the concepts or vocabulary to fully describe bat phenomenology should not lead us to dismiss the reality of their experiences, and our inability to describe or understand something does not negate its existence. "What is it like to be an AI?", or if there is even something that is like, is an issue we continue exploring through Murray Shanahan's 20 questions experiment, explained in the section 4.2.5.3 of this thesis about "Signifying Uncertainty".

11. The word "usually" is intentionally used in order to be inclusive with the blind and deaf communities, and recognize the feat of blind people who trained themselves to use a human analogue of echolocation to sense space, for a fascinating introductory account, one can consult the Daniel Kish's TED Talk "How I use **sonar** to navigate the world" [23], whose subtitles the author of this thesis actually translated to Catalan.

allows us to think abstractly and be aware of our own experience and environment in a very different way than either animal.

Another intersection point between Schramm and Uexküll is the recognition of how deeply our environment shapes our cognitive development, and the background of our everyday thinking and understanding. He illustrates this formative environmental influence with a beautiful metaphor of stalagmite formation:

> *[...] a message is much more likely to succeed if it fits the patterns of understandings, attitudes, values, and goals that a receiver has; or, at least, if it starts with this pattern and tries to reshape it slightly. Communication-research men call this latter process "canalizing," meaning that the sender provides a channel to direct the already existing motives in the receiver. Advertising men and propagandists say it more bluntly; they say that a communicator must "start where the audience is." Teachers know they should start "where the pupils are. You can see their personalities -our patterns of habits, attitudes, drives, values, and so forth- grow very slowly but firmly. I have elsewhere compared this channeling process to the slow, sure, ponderous growth of a* **stalagmite on a cave floor**. *The stalagmite builds up from the calcareous residue of the water dripping on it from the cave roof. Each drop leaves only a tiny residue, and it is very seldom that we can detect the residue of any single drop, or that any single drop will make a fundamental change in the shape or appearance of the stalagmite. Yet, together, all these drops do build the stalagmite, and over the years it changes considerably in size and somewhat in shape. This is the way our environment drips into us, drop by drop, each drop leaving a little residue, each tending to follow the existing pattern [18].*

In the metaphor, each individual experience or interaction may leave only a minimal trace, often imperceptible, yet collectively, these countless small influences gradually accumulate, building and reshaping our mental models over time, following existing patterns while growing and subtly altering them. This process of "**environmental dripping**" constructs our mental models incrementally, and is consistent with modern cognitive science paradigms such as embedded cognition (and potentially, predictive processing).

What all of this evinces for human-AI interaction is how paramount it is to connect experiential fields. First, **designers and developers** must have awareness and empathy for users' experiences and mental models. Second, **users** need accessible conceptual models of AI capabilities and limitations. Third, **CAs** require useful approximations of human needs and intentions. When any of these bridges collapse,

when there is not sufficient overlap, communication is prone to failure. A user's linguistically perfect prompt accomplishes nothing if it references contexts beyond the AI's training or access, or attaches files it cannot read properly. Similarly, powerful AI capabilities remain inaccessible if interfaces and documentation developed by the industry make them opaque, neglecting the public imaginary and users' existing or absent mental models for interacting with CAs. Communication thus becomes one of the weakest links in the chain of human-AI interaction on both of the following accounts: Humans and artificial agents inhabit fundamentally different experiential worlds, while AI experts and most everyday users operate from drastically different fields of experience. The next chapter in this thesis (chapter 4) will explore the know-how of interface design, visual language and the psychology behind the use of everyday things or its failure.

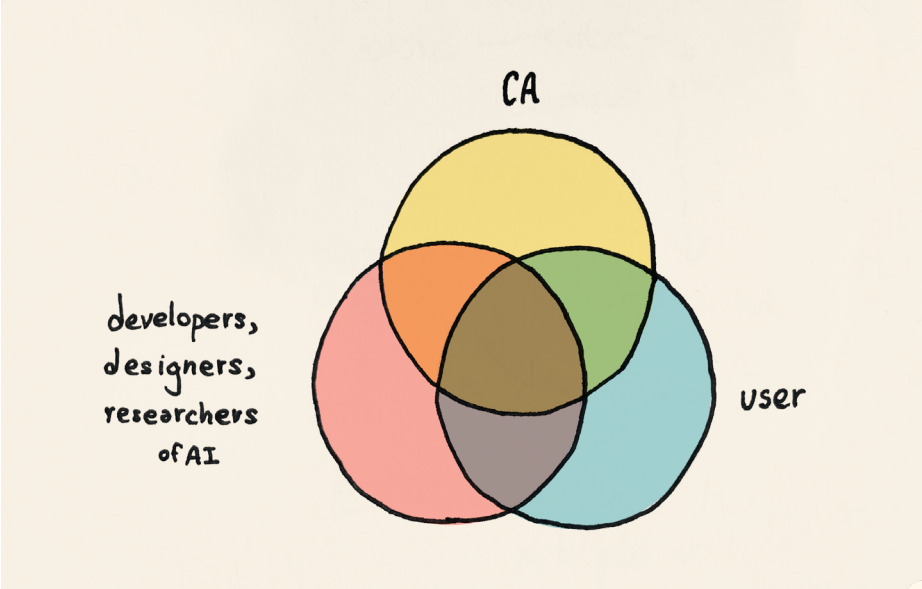


Figure 11. *Overlap between the fields of experience of CAs, everyday users, and developers, designers and researchers of AI.* Effective human-AI communication depends on sufficient overlap across all three: those who research, develop and design CAs must find ways to make their understanding accessible to users, users need functional models of CAs, and CAs require approximations of human needs and context. Application of Schramm's overlapping fields of experience to human-AI interaction. Original illustration by the author.

⚑ **Overlaps of experience**

Field of experience: totality of what a communicator has lived through, learned, and internalized (knowledge, beliefs, values, cultural background).

- They are shaped incrementally, and **environmentally** (Schramm's stalagmite metaphor).

→ Communication depends on the degree of **overlap between communicators' fields of experience**.

For human-AI interaction:

1. Between **humans and CAs**, whose experiential worlds differ structurally (Uexküll's Umwelts).
2. Between **AI practitioners and everyday users**, whose expertise of the same system differs drastically.

3.2.2.2 Feedback and Interpretation: From Pipes to Steaks

In what follows, we will provide an overview of how **(1)** feedback and **(2)** interpretation transform communication in Schramm's model, and to illustrate these more clearly **(3)** we will bring in several examples from different sources. And finally, as promised in the previous sub-chapter, **(4)** we will look into the concepts of feedback and bi-directionality in more detail.

(1) Schramm's model also incorporates the concept of **feedback** as the return process in communication, where the receiver communicates back to the sender. This provides information about how the message is being interpreted, allowing the sender to adjust their communication. Imagine a conversation with a friend. If you are trying to persuade your friend, feedback might return in the form of them saying "This makes sense" (verbal responses), nodding in agreement or frowning in disbelief (nonverbal cues). These reactions are all ways the receiver is communicating back their mental state, indicating their understanding or (dis)agreement. Similar examples at the level of mass communication include a letter to the editor protesting or congratulating an editorial, or even the applause of a lecturer's audience [18].

This notion fundamentally transforms communication from a one-way linear transmission into something dynamic and bidirectional. This apparently simple addition to communication models, symbolized by one more message and arrow in a diagram, changes the foundations of communication systems by introducing **circularity** and temporal dimensions that enable continuous refinement and adaptation. Without feedback, communication is essentially a shot in the darkness. With feedback, it allows communicators to assess how their messages have been interpreted, to adjust subsequent communications accordingly, and thus, becomes a collaborative process where meaning is negotiated and refined until shared understanding is achieved.

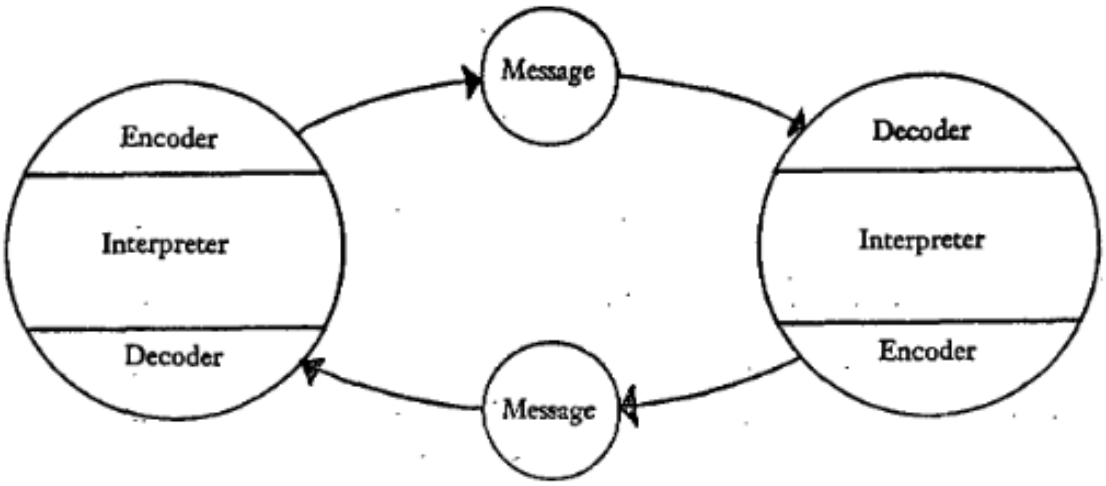


Figure 12. *Schramm's circular model of communication.* It illustrates communication as a dynamic and reciprocal process, where each participant simultaneously performs the functions of encoding, interpreting, and decoding messages. Original from Schramm (1954).

(2) Following the introduction of feedback, it is important to understand how a receiver actively makes sense of a message. For Schramm, this step is carried out by the **interpreter**. Imagine you are listening to a friend tell a story, and as they speak, you are not just hearing sounds; your brain is actively decoding those sounds into words and then going further to interpret what those words mean based on your own past experiences and your friend's, your understanding of language, and the context of the conversation. This means what Schramm calls the "picture in your head" might be slightly different from the "picture in your friend's head", because your fields of experience will influence how you assign meaning to their words. Essentially, the interpreter is the active decoder (on the receiving side) who takes the signs of a message and builds their own understanding of what the sender is trying to communicate. For Schramm, this concept is needed to explain how communi-

cation organizations like newspapers, broadcasting stations, and film studios act as institutionalized interpreters. These organizations actively decode inputs like news reports, interpret their significance, and then encode them into messages for a wider audience. Therefore, interpretation, whether at the individual or organizational level, is the active process of taking in signs and making sense of them by drawing upon experience, current state, and the context in which the signs are received.

Now, let us take a step back and remember that Shannon and Weaver make a clear distinction between the message and the signal (transmitted) [13] with the implications we saw in the discussion of their model. Schramm goes further and takes a different approach, he asks "What happens when a signal comes to you?" and reminds us that it comes in the form of a **sign**, and that if we have learned the sign, we have learned certain responses with it.

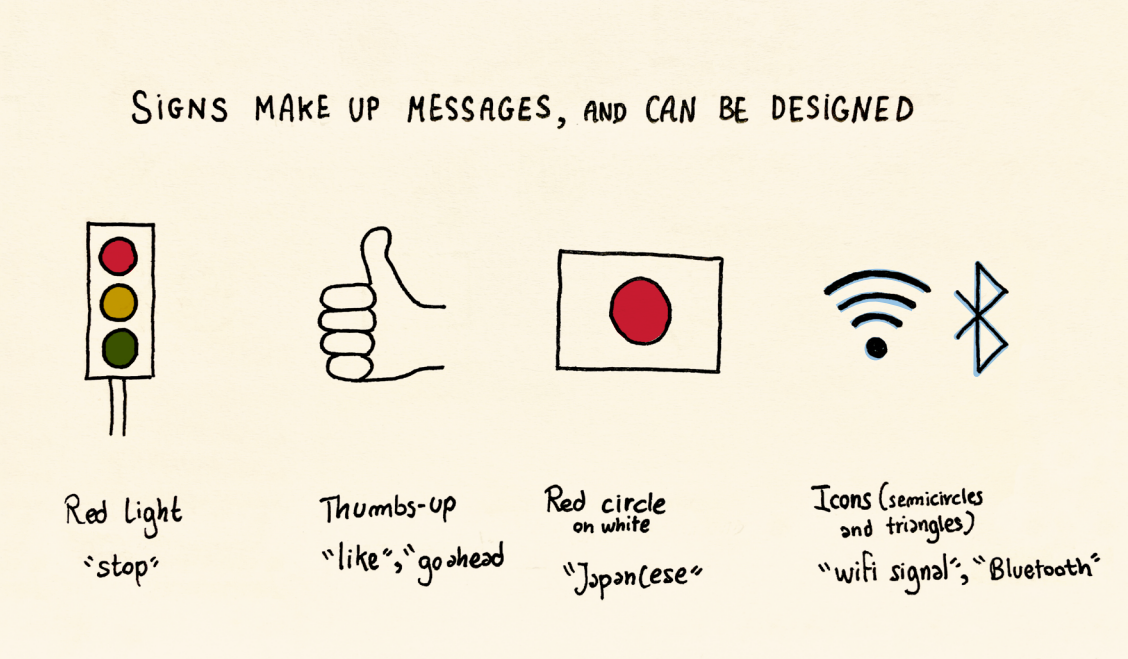


Figure 13. *Four examples of signs in everyday life*: a red traffic light, a thumbs-up gesture, a national flag, and geometric icons for invisible phenomena like Wi-Fi and Bluetooth. By introducing signs, Schramm implicitly makes communication something that can be deliberately designed. Original illustration by the author.

(3) Let us take an example to visualize the ubiquity of signs in our everyday life; think of traffic lights, in which a red light triggers us to stop, hand gestures such as a thumbs-up or flags. In our GUIs, signs are everywhere, from the symbols for Wi-fi or Bluetooth, essentially standing for something invisible to sight, to the hamburg-

er menu standard icon (three horizontal lines, ☰), signing that there are multiple menu options available within. These digital signs function like traffic lights or hand gestures in the physical world, triggering learned responses if users interpret their intended meanings. The effectiveness of an interface often depends on how intuitively users can decode these signs without explicit instruction, which makes interface design a thoughtful exercise in creating and arranging signs.

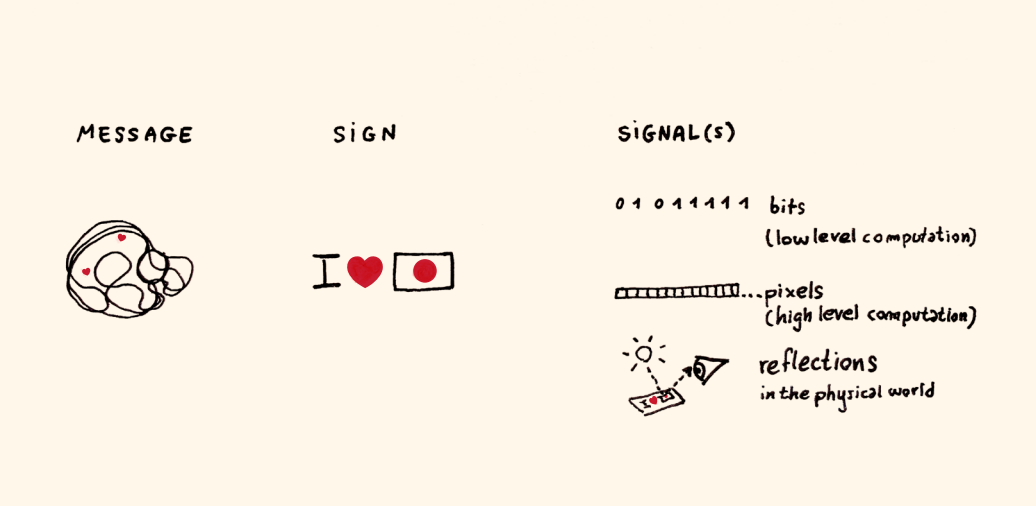


Figure 14. The thought or feeling to be communicated is encoded into the signs of a message, which can in turn be transmitted through multiple nested layers. Each element (bits[12], pixels, light) might function as a signal at one level, but be the message that a different transmitter-receiver encodes, transmits and decodes through a different signal at another level. This connects Schramm's semiotic perspective with the layered transmission described by Shannon and Weaver. Original illustration by the author.

By introducing signs, Schramm implicitly creates a conceptual bridge that connects to Weaver's three levels of communication problems. For clarity and analytical precision, we can distinguish between: **Signal (Level A - technical) ≠ Sign (Level B - semantic) ≠ Message (Level C - effectiveness**, intended meaning and desired effect**)**. It is noteworthy that a single message (such as "you have to stop") can be alternatively represented by multiple signs (such as a red traffic light, and a red octagonal "STOP"). Encoding a message might also require multiple signs, and signs of multiple kinds, such as expressing "I love Japan" with "I(Heart)(Japanese flag)" (see Figure 14 below), which uses a combination of linguistic symbols (the word "I",

12. If bits can be considered signs at the machine level, processed by the system into values such as pixel colors, is a question that connects to Shannon's formal definition of information (see footnote 5).

standing for oneself) and visual symbols (the heart and the flag, standing for "Love" and "Japan"). René Magritte's surrealist painting "The Treachery of Images" (1929) offers a perfect visualization of this distinction with an illustration of a pipe with the subtitle "Ceci n'est pas une pipe", that questions the conventions of both verbal language and visual imagery.

Moreover, the nested layers of communication we elaborated on in the discussion of Shannon-Weaver's model entails a chain of translations. After translating the message to signs, it is sequentially transmitted as a signal multiple times, from the pixels on a screen (high level computational representations) to a vector of bits (low-level computational) before being ultimately converted to reflected light stimulating the rods and cones in our retinas, ultimately becoming neural signals that we will interpret. This multi-level translation process highlights why communication across the human-AI interaction is complex, with not only signal degradation and constraints, but also **noise** (of the semantic, effective and conceptual model kind) **across these translations**.

Schramm also brings to the picture and highlights the role of **context and internal** (mental) states in interpretation, and uses as an example a picture of a steak, which "may not arouse exactly the same response in you [if you are hungry] as when you are overfed". This context-dependent interpretation aligns with embodied cognition theories in cognitive science, which posit that our understanding is fundamentally shaped by our bodily states and environmental contexts. Its implications for interface design are that user interpretations will shift based on factors like fatigue, stress, or environmental distractions. The same interface elements might be processed differently by the same user under different conditions, and robust designs must account not only for a diversity of users and fields of experience, but also for a vast range of interpretive contexts.

Finally, among what comes into play in this interpretation, Schramm states we are constantly decoding signs from the environment (as we pointed out with the **environmental dripping** a few paragraphs ago), interpreting these signs, and encoding something as a result. He believes that:

> *In fact, it is misleading to think of the communication process as starting somewhere and ending somewhere. It is really endless. We are little switchboard centers handling and rerouting the great endless* **current** *of communication. We can accurately think of communication as passing through us, changed, to be sure, by our interpretations, our*

> *habits, our abilities and capabilities, but the input still being reflected in the output [18].*

In other words, the environment would be coupled and constantly shaping our cognitive processes, definitely a **mediator**, and arguably even a constituent. Our interpretations change what we experience, thus **shaping** the fields of experience that will in turn shape how we live in and interpret our environment next, and all of our interactions with everything in it. This line of thinking is very much consistent with 4E cognition frameworks (embedded, extended, embodied and enacted cognition).

(4) This brings us to bidirectionality introduced by feedback which makes each communicator alternate between the roles of encoder and decoder. Furthermore, it allows to adjust further messaging, for example, accounting for the noises, and, in following Schramm's formulation, whether the "picture in the head" of the receiver will bear any resemblance to that in the head of the sender [18]. The notion of feedback as a **self-regulatory mechanism** and the circularity it introduces are inextricably tied to some of his contemporaries and predecessors, from James Clerk Maxwell's early paper "On governors" (1868) [24] to cybernetics. The term cybernetics was proposed by Norbert Wiener in his foundational work "Cybernetics or Control and Communication in the Animal and the Machine" (1948) as study of "a set of problems centering about communication, control and statistical mechanics, whether in the machine or in the living tissue" [25]. This interdisciplinary approach emphasized precisely that communication flows in a circle rather than in a line, feedback being the key element that turns this line into a circle. In other words, both Schramm's model of communication and cybernetics reject linear causality in favor of circular patterns where outputs become inputs for the other cycles, and communication becomes an endless reciprocal flow. This introduces an increased[13] (but necessary) **complexity** for interaction, and makes what we discussed in Shannon-Weaver's model as nested layers of communication potentially scale into a series of inter-related feedback loops.

That when we receive a puzzled look back from the person to whom we are speaking, or a question, we spontaneously adjust our explanations, is in a way a cybernetic control mechanism operating through social interaction. In human-AI interaction, feedback becomes even more crucial due to the fundamentally different experience

13. While Shannon includes the concept of a "**correcting device**" in the context of noisy channels where an observer transmits "correction data", that is thought by Shannon of as a separate element, rather than something within the same loop, and the term "feedback" is not explicitly mentioned as a core element of the communication system.

fields of the human and the CA. Without robust feedback channels, a user cannot know if their instructions have been correctly interpreted, and the CA remains unaware of misalignments between its responses and user intentions. The frequent frustrations in human-AI communication often stem from **feedback channels** that are too narrow, delayed, misused, underestimated, or entirely absent in our interfaces.

For effective CA interfaces, we need to prioritize **feedback mechanisms** that make some degree of explainability possible (without being overwhelming), and interpretation visible, so communication can be adjusted by both parties. This could manifest in simple ways for the users, such as an interactive tutorial in the GUI progressively disclosing and assessing understanding for beginners (but with options to skip and access it back later), or available ways to sign they might require help, and to get it back, such as displaying tooltips (little box or popup) when we hover over a GUI element (such as a button or an icon), for more than half a second. For the CAs, it can be explicit verbal requests to confirm and clarify complex user instructions, or their ambiguous language (Level B) and intentions (Level C). They can also manifest in more elaborate ways, as we will explore in chapter 4, in which we treat with more depth the concept of feedback with Don Norman and the bases of interaction design and how it relates to XAI. Both with simple solutions and complex endeavours, it is through communication and constant feedback that both us and artificial communicators can continuously update our conceptual models of each other, creating the "commonness" or overlap that Schramm identifies as the necessary core for successful communication.

⚑ Signs, feedback, and interpretation

Feedback transforms communication from linear transmission into a circular, self-regulating loop, in which expectations are continuously adjusted.

Sign: basic unit of meaning in communication (Level B). A message can be represented by one or multiple signs. The same sign can be **interpreted** differently by each communicator, or even the same one at different points (depending on previous experience, context, and internal states).

→ In interface design, design choices (an anthropomorphic avatar, a typing indicator, a Google-like layout) are **interpretable signs**, and can predispose certain communication.

3.2.3 Limitations of Schramm's Model of Communication

Schramm's model is born with the seeds of the media studies of the 1950s and shifts the war-born perspective of communication as mechanical transmission to **collaborative meaning-making**. Shannon's perspective on communication as signal transfer, then expanded by Weaver, transforms with Schramm, who looks at communication (consistently with its etymology) as establishing **"commonness"** (from Latin "communis") between the parties of communication.

As is the case with any model, Schramm's model too retains important limitations for our context. As John von Neumann put it in "The Mathematician" (1947), truth is "much too complicated to allow anything but approximations" [26]. We can summarize these limitations as follows: **(1)** While Schramm's model advances beyond linear transmission, it still conceptualizes communication as **sequential turn-taking** rather than the simultaneous encoding-decoding that characterizes complex human-AI interactions. **(2)** The model's treatment of **experience fields** remains relatively simplified, not fully accounting for the nested layers of communication we identified, nor the fundamental asymmetry between human phenomenological experience and that of AI.

(1) The distinction between linear transmission models like Shannon-Weaver's and interaction models exemplified by Schramm's work represents a fundamental advancement in communication theory. While linear models conceptualize communication as unidirectional information transfer without confirmation of receipt, interaction models recognize the bidirectional nature of communication where participants dynamically alternate between sender and receiver roles. This shift introduces feedback as a crucial mechanism that enables message verification, comprehension assessment, and adaptive adjustment of communication strategies. Interactive communication naturally introduces greater complexity through turn-taking sequences, such as in Q&A sessions, and instant messaging conversations where participants send messages and await responses. Though this interactivity adds necessary complexity to our theoretical framework, it more accurately captures the collaborative, recursive nature of communication processes with current CAs.

While Schramm's model significantly advances our understanding of communication, it is still concerned with communication as a sequential process of alternating turns. **Transactional models** like Dean Barnlund's (1970) address this limitation by recognizing that sending and responding occur **simultaneously** rather than sequentially.

Furthermore, Barnlund aims for his model to give an account of communication that encompasses both its interpersonal and its intrapersonal levels, and frames communication as "the production of meaning, rather than the production of messages," highlighting that meaning emerges dynamically.

Another limitation of Schramm's model is its relative simplification about what happens within each communicative agent. While Schramm acknowledges internal processes, **intrapersonal models** like Barker and Wiseman's see them in far more detail, adding concepts such as **internal stimuli** and **internal self-feedback**. These complementary approaches acknowledge that no single model fully captures the complexity of real interaction, but each contributes valuable insights that, when synthesized thoughtfully, provide a more comprehensive foundation for understanding, analyzing and designing these communication systems.

(2) Schramm's model introduces that an **overlap in the fields of experience** of the parties is a requirement for effective communication between participants, and it aligns with Weaver's insights that effective communication does not only depend on signal fidelity. When it comes to human-AI interaction, humans bring experiential fields built from embodied cognition such as the weight of objects, the warmth of sunlight, the frustration of miscommunication, or the rewarding sense of understanding. However, the "experience" of the CAs is constructed entirely from vicarious **linguistic relations** (millions of verbal descriptions about weight, warmth, frustration and reward). In some ways, it is like trying to establish commonness between someone who has lived in a house and someone who has only read architectural blueprints. They might both use the word "home," but their experiential fields barely overlap, and their interpretations will differ. Human experiential stalagmites form through embodied and environmental dripping, but CAs can only construct their representations through the verbal descriptions of environments they never directly interact with due to the interactions allowed through their interfaces.

3.2.4 Synthesis

In this chapter we have dealt in depth with the communication models that we deemed more complete and explanatory for our particular object of study (human-AI interaction) among those that were foundational for the fields of information theory and media theory. We have examined in depth the relevant aspects of a representa-

tive of **linear transmission models** (Shannon-Weaver) and a representative of **interactional models** (Schramm). This decision has been motivated by the **hybridity** of our object of study; **human-AI interaction**. While linear transmission models were conceived thinking of machines and artificial systems in mind, interactional models were conceived for human interaction. There are tradeoffs in either approach, but we believe that it is by combining what we can learn from both that we can understand human-AI interaction.

In this chapter, we mainly focused on communication models, which were developed to understand message transmission across a channel, encoding and decoding processes, noise and fidelity, feedback loops, context and shared frames of reference. Human-AI interaction involves all of these, and instead of reinventing the wheel, we can use the conceptual tools of communication models for analyzing it. The **Shannon-Weaver model** provides ways to systematically think about where **information is lost or distorted**, and **Schramm's model** integrates subjective fields of experience and **feedback** that shape how meaning is interpreted and co-constructed in bidirectional interaction. Weaver's and Schramm's perspectives bring into communication theory the ambiguity and relativity of language for conversational agents, the relevance of the environment, the interlocutors, and fields of experience that resemble Uexküll's umwelts. The latter evince how an **overlap between fields of experience** is necessary for communication. This gives way to the distinction between the signal (Level A - technical), the sign (Level B - semantic) and the message (Level C - effectiveness).

From what we have seen up to this point, we have sufficient content to provide a more robust answer to our initial questions on if interfaces are communicative artifacts, and if interface design is a communicative act. We have also broken down what happens in a communication system and how does the meaning of a message persist or transform. What now remains to answer is "where do interfaces sit in this system?" Here we synthesize these theoretical insights into a diagram that illustrates how interfaces both function as medium, messages themselves, or transformative interpreters (CA-user communication and designer-to-user communications).

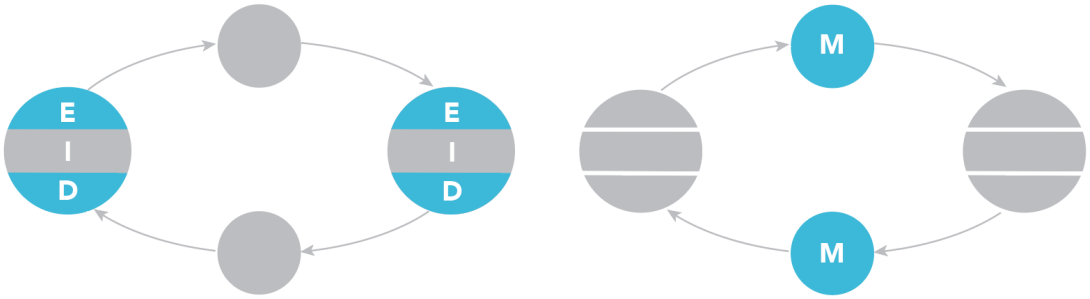


Figure 15. *Schramm's communication model with interpretation and feedback, adapted to visualize the role of the interface (in blue).* Left, the interface as a decoder and encoder mediating between communicators. Right, the interface as a message, from AI practitioners to users. By making the interface visible within the model, we can recognize interface design as a communicative act. Original illustration by the author, adapted from Schramm (1954).

By mapping these relationships visually, we can more precisely identify the points where meaning persists or transforms as it travels, gets lost and reconstructed in a communicative system between humans and CAs. This visual approach has provided another piece to the intellectual puzzle we would like to draw, contributing to answer for the role of interfaces for understanding interface design as both a communicative act and a bridge between fundamentally different experiential worlds.

An explicit focus on the concept of interface is lacking in the models discussed in this chapter. Because our topic is human-AI interaction, this concept comes to occupy the central stage and an interface becomes not only an encoder and decoder for a signal, but also a **message** itself. At this level of analysis, the source of the message are **AI developers, designers and practitioners**, and the destination are their **clients**, the users. Therefore we need an explicit focus on design choices, such as whether to use an anthropomorphic avatar or a Google-like layout, or if showing a "…" typing indicator that resembles social messaging apps, which are functional cognitive choices as much as aesthetic ones; about what information is transmitted in which channel, how shared context is established and maintained, what feedback loops are enabled, and how misunderstandings are detected and repaired.

All of this makes of **interface design** not only an aesthetic or usability concern, but a necessary and rather underused component of human-AI interaction, and AI functionality. The interface becomes the main communication medium through which humans and AI systems develop effective communication, and a powerful message

shaping how users interpret what to do with CAs.

Design which can be seen through the lens of communication theory as a **communicative act** will be the focus in the next chapter. The **interfaces of CAs** are a **communicative artifact**. For example, graphic user interfaces (GUIs) do not only display written messages, but through their layouts, styles, elements and design choices can provide useful visual cues that guide user interaction with conversational agents (CAs).

⚑ Where do interfaces sit in a communication system?

The most relevant source-destination pairs for human-AI interaction:

1. **AI practitioners and designers → everyday users.**
2. **Conversational agent ↔ humans.**

The interface is simultaneously a medium (**encoding and decoding** between user and CA) and a **message** (from AI practitioners to users, communicating what the CA is and how to interact with it).

From the lens of communication theory, interface design is a **communicative act**, and the interfaces of CAs are **communicative artifacts**.

4. *How.* The Language of Interaction Design

An artifact is basically a thing that is made, from Latin "arte" ("by or using art") and "factum" ("something made"). The previous chapter established that interfaces are communicative artifacts: they transmit messages but also are messages themselves, conveying decisions and guidance from practitioners to users. This chapter turns from the "why" of that communication to the "how" and the principles through which interfaces address it.

Connecting artifacts to human understanding has been a persistent design concern, embedded in the very origins of the discipline[14]. However, the theoretical work for undertaking the endeavor of thinking about how artifacts make sense to people is fragmented across disciplines, from cognitive sciences to engineering, fine arts and media studies.

Herbert Simon proposed artifacts function precisely as "interfaces" between an "inner environment" (substance and organization of the artifact itself), and an "outer environment" (which represents the surroundings in which it operates) and presented design as the process of "**adapting** the inner **environment** to the outer environment to attain specific goals", concerned with how things ought to be (creative synthesis often characteristic of architecture, engineering, or medicine), instead of how things are (and the descriptive efforts often characteristic of natural sciences) [27]. Klaus Krippendorff directly asks whether there could be "a **science for design**" [28]. Richard Buchanan echoes it, stating "the significance of seeking a scientific basis for design" lies in a "concern to connect and **integrate** useful knowledge from the arts and sciences alike", and characterizes through Horst Rittel design problems as "**wicked**" (indeterminate, with many interdependent factors, often incomplete or dynamic). These thinkers point out that designing artifacts which make sense to their users requires interdisciplinary perspectives and systematic thinking about human cognition [29] [30].

Consistently with them, the perspective of this thesis looks at **graphic user interfaces** as **structured communication** that can engage our visual thinking capacities by creating cognitive scaffolding that mediates understanding between human and

14. Even in its Latin etymology, the word design itself carries communicative intent too, combining "de" ("out") with "signare" ("to mark"), from "signum" ("identifying mark, sign"). In this we find the roots of design as both a creative activity and deliberate making.

artificial systems (rather than decorative or arbitrary collections of elements)[15]. The author whose work we will discuss next is distinctive in this landscape: developed jointly from engineering and cognitive psychology, it addresses precisely the questions around how users form and update conceptual models of systems through interaction. It also provides the cognitive vocabulary (such as affordances, signifiers, conceptual models) required for a systematic analysis of the interaction between humans and CAs, necessary to analyze the challenges of interface communication.

4.1 Norman's Design Principles

In this chapter, we will focus on Don Norman's Design Principles. Let us start with an everyday example. When was the last time we struggled with a door? Sometimes we push instead of pulling, or search for a handle that is not there. That we struggle more often than necessary with the design of rather simple mechanisms in our daily lives shows how human cognition interacts with designed environments. When we approach a door, we bring unconscious expectations about how it should work based on previous experiences and current perception. Door designs that work against our intuitive understanding bring confusion and force us to redirect cognitive effort away from our goals to consciously think about actions that would otherwise be automatic. These interactions between humans and designed objects (artifacts) relate to core cognitive science concepts: mental models (our internal representations of how things work), active inference (how we predict outcomes based on prior experience and current perception), cognitive load (the mental effort required for the task in hand, in our working memory), and 4E cognition approaches (how our understanding is shaped by our physical capabilities and interactions with the environment) such as enacted, embedded, embodied, and extended cognition.

If one has ever had problems with light switches in an unfamiliar room or struggled with foreign water faucets, coffee machines and other home appliances, they have experienced the origin point of what we are about to discuss. Don Norman's "The Design of Everyday Things" (originally "The Psychology of Everyday Things", 1988) was inspired exactly by these everyday frustrations when he moved from the US to the UK for a sabbatical in the Applied Psychology Unit in Cambridge [31] [32]. His

15. A stance consistent with embodied and situated cognition approaches (within 4E cognition), which emphasize how cognitive processes engage with external structures in the environment.

personal frustration with everyday objects, treated with insight and curiosity, and thorough discussions with colleagues, were the starting point for thoughtful research that later drove innovations to reshape interaction and a more human-centered design. To start, Norman posed a deceptively simple question: When we first encounter something we have never seen before, how do we know what to do? This question, which appears deceptively simple for physical objects like doors or teapots, becomes even more complex and relevant when applied to systems with no physical form to communicate their capabilities and limitations.

For Norman, "all **artificial things are designed**", but not all of them entail physical mechanisms. This includes services and procedures, from teaching to business and governance. Design is concerned with "how things work, how they are controlled". The focus is placed on the "interplay between technology and people to ensure that the products actually fulfill human needs while being understandable and usable" and extends design to the making of the "delightful and enjoyable, which means that not only must the requirements of engineering, manufacturing, and ergonomics be satisfied, but attention must be paid to the entire experience, which means the aesthetics of form and the quality of interaction" [31]. He emphasizes the importance that "when done well, the results [of design] are brilliant, pleasurable products", often effectively invisible and intuitive, but rather noticeably "when done badly, the products are unusable, leading to great frustration and irritation." In his conception, he integrates industrial design, interaction design and experience design, placing human cognition and the effect and relationship with human-made objects or technologies at the center of the design process in a way much consistent with his later work [32] [33]. In other words, design would be the process of making the artificial understandable, usable, and preferably, enjoyable.

While Norman advocated for User-Centered Design (UCD) in human-computer interaction as early as 1986 [34], in later work he explicitly redefined it as **Human-Centered Design (HCD)**, conceiving the user as a broader person ("an approach that puts human needs, capabilities, and behavior first, then designs to accommodate those needs, capabilities, and ways of behaving" [31]). In his latest book and endeavours, he has broadened the scope to advocates for Humanity-Centered Design (HCD+), shifting the focus from single separated individuals to their relation with their world, history and sustainability, and encouraging the industry to be conscious about the broader societal impact of technologies [35] [36].

4.1.1 Discoverability and Understanding

According to Don Norman, two of the most important characteristics of good design are discoverability and understanding.

Discoverability addresses the question: "Is it possible to even figure out what actions are possible and where and how to perform them?" [31]. This focuses on ensuring users can identify potential interactions within a system, and kick-start the interaction intuitively.

- We might think of that time in which we could not find the button to turn something on (perhaps because there was none, such as in bathrooms with motion sensors instead of physical light switches on the wall); we knew the possibility of turning the light was there, but not where and how to activate it. Or that other time in which we discovered a relevant feature of a device or a service months or years after starting to use it (such as a keyboard shortcut activating something useful in a desktop app, or a swipe for deleting or double-tap for sending a like in a smartphone app); we did not even know the possibility was there. The fanciest features of the best systems are of little use if the users do not discover they are available to them.

Understanding addresses the questions: "What does it all mean? How is the product supposed to be used? What do all the different controls and settings mean?" [31]. This more semantic dimension of design focuses on ensuring users can interpret a product's intended functionality, make sense of its interactive elements, and conceptualize its operation.

- We might think of that time in which we stared at a common home appliance or car dashboard and wondered if a spaceship control room would not be clearer (examples include washing machines with dials and labels such as "Normal", "Program 1", or "Special Care" that do not clarify what types of clothes they are for, nor temperature, spin speed or water usage). Or that other time in which we got a vague technical error message so helpfully vague that it did not explain what went wrong or how to fix it (such as "Your request could not be processed", "Something went wrong" or "Connection Error").

When combined, the **qualities of being discoverable and understandable** help ensure that products are both **accessible and meaningful** to the people who use them. Norman states that "**discoverability** results from appropriate application of

five fundamental psychological concepts: **affordances, signifiers, constraints, mappings, and feedback**," to which we will go in depth in the following paragraph, "but there is a sixth principle, perhaps most important of all: the **conceptual model** of the system. It is the conceptual model that provides true understanding."

Together, these 7 concepts[16] establish a way to conceptualize how users navigate and engage with designed environments, highlighting Norman's central idea that effective design must depart from the knowledge basis of human cognition in order to facilitate both the detection of possibilities and meaningful interpretation of whatever is designed. In the following section, we will use as a guideline to his work what Norman calls Fundamental Principles of Interaction, derived from understanding how people behave and think in context. These principles include:

- Five fundamental psychological concepts applied to the designed objects: affordances, signifiers, mappings, constraints, and feedback.
- The conceptual model of the system, which as he emphasizes, is perhaps the most important, for its relation to genuine understanding, and is thus included as one of the principles. As designers, we develop our own conceptual understanding of a system and, through our design choices and documentation, inevitably influence which mental models users are likely to form. However, the ultimate construction and adoption of these mental models remains within the cognitive domain of the users themselves.
- Discoverability, the meta-principle we just discussed, which emerges when the other principles are properly implemented.

16. While in Norman's first chapter of "The Design of Everyday Things", these concepts are called "fundamental principles of interaction", and mainly 5 are introduced in length (affordances, signifiers, mapping, feedback, conceptual models), while another (discoverability) is mentioned, in the second chapter he calls them "fundamental design principles", making a parallel with seven stages of action and aggregating 7 principles: adding discoverability and constraints to the 5 psychological concepts. They are reordered as: discoverability, feedback, conceptual model, affordances, signifiers, mappings, constraints. The explanations in this thesis incorporate also plenty of what he analyses in later chapters of the book, such as kinds of constraints and feedback, or its relation with human error. We mention this because these principles are presented as 5, 6 or 7 in other sources, including Norman's first edition of the book. For our explanations, we have taken as a departure point the latest edition of the book (2013), and our order is inspired in the initial order starting with affordances, signifiers, mappings, and feedback. Then we add constraints and conceptual models, which closes the circle started by discoverability and understanding, that we introduced first to provide a revealing snippet of Norman's contextual basis and perspective.

These principles, while presented separately, function in an interconnected way, and what they describe is the base of our interaction with designed objects and systems.

4.1.2 Affordances (and Possibilities)

The concept of affordance addresses a challenging but paramount question in both cognitive science and design: how do we perceive what actions are possible in our environment?

While the term was introduced to the design community by Don Norman, he warmly credits his colleague, the perceptual psychologist James J. Gibson, who introduced it in "The Senses Considered as Perceptual Systems" (1966) and more thoroughly developed it in "The Ecological Approach to Visual Perception" (1979) [37] [38]. To Gibson, affordances are relationships that exist naturally, which means they do not have to be visible, known, or desirable [32]. They are simply the **action possibilities** that exist in an environment [33]. The relational nature of affordances challenges our typical thinking about objects and their properties. But "affordance is not a property" of the object itself; whether an affordance exists depends on the properties of both the object and the agent [31] [32]. In Norman's definition and example:

> *The term affordance refers to the* **relationship between** *a* **physical object** *and a person (or for that matter, any* **interacting agent***, whether animal or human, or even machines and robots). An affordance is a relationship between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used. A* **chair** *affords ("is for") support and, therefore, affords sitting. Most chairs can also be carried by a single person (they afford lifting), but some can only be lifted by a strong person or by a team of people. If young or relatively weak people cannot lift a chair, then for these people, the chair does not have that affordance, it does not afford lifting. The presence of an affordance is jointly determined by the qualities of the object and the abilities of the agent that is interacting [31].*

When Norman asks us to imagine a chair, it affords sitting because of the relationship between its physical structure and the ability of our human bodies to bend at specific joints. However, the same chair might not afford sitting for a different species, such as ants or octopuses, or for those with different physical capabilities. The key to this elusive concept is that the physical properties of the object **allow or enable these actions**, but **for** a given actor.

As we see it, the distinction between Norman's and Gibson's view of affordances lies on the **debate** between **mediation and constitution** of cognition that we mentioned in the introduction. Gibson's view is rooted in his ecological approach to perception. He believed that perception involves the direct "pickup of information" from the environment [32] [31] [38]. Norman shows a practical concern with user interpretation and usability, putting the emphasis on making affordances discoverable, which can be seen as leaning towards the idea of mediation, where design features mediate the user's understanding of what actions are possible. While they agreed on most points regarding the definition of affordance and they see eye to eye in that "the required information is in the world", Norman assesses they diverged in interpreting "how the mind actually processes perceptual information", regardless of their innumerable and friendly intellectual discussions [32].

When Norman introduced it, the term affordance caught on with overwhelming popularity amidst the design and later human-computer interaction research communities [39] [40] [41]. However, it led to quite a bit of confusion, with what Norman calls some professionals saying things like "we need to put an affordance on the UI". Norman, concerned about the misuse, has earnestly tried to **clarify** the controversy of its definition through his later work distinguishing between real and perceived affordances [32], repeatedly specifying he was referring to perceivable affordances in his earlier work, introducing the concept of signifiers [33], and making a thorough revision of the whole book in which he presented them [31]. In fact, Norman even asked to "forget affordances: what people need, and what design must provide, are signifiers" [33].

> *Consider the traditional computer screen where the user can move the cursor to any location on the screen and click the mouse button at anytime. In this circumstance, designers sometimes will say that when they put an icon, cursor, or other target on the screen, they have* **added an "affordance"** *to the system. This is a* **misuse of the concept**. *The affordance exists independently of what is visible on the screen. Those displays are not affordances; they are visual feedback that advertise the affordances. […] Similarly, it is wrong to claim that the design of a graphical object on the screen* **"affords clicking."** *Sure, you can click on the object, but you can click anywhere. Yes, the object provides a target and it helps the user know where to click and maybe even what to expect in return, but those aren't affordances [32].*

A significant distinction is between "**real** affordances" and "**perceived** affordances."

Real affordances are the actual possibilities for action, while perceived affordances are what we perceive as possible. In physical product design, both types exist simultaneously. In **digital devices**, the only real affordances are the ones that come with the device, such as clicking with a mouse, typing on a keyboard, displaying information on a screen or tapping on a smartphone. Therefore, in **GUIs** and screen-based graphic design, we control only perceived affordances. A good way to visualize this is that "although all screens within reaching distance afford touching, only some can detect the touch and respond to it" [32]. And even when "a display would afford pointing, touching, looking and clicking on every pixel of the screen" [32], one of the aims of human-centered design is to create and use metaphors and conventions in order to provide the right signifiers, mapping, constraints and feedback to make affordances discoverable and perceivable to people and guide interaction.

Finally, to round the definition of affordances, it is worth noticing that affordances have a counterpart: **anti-affordances**. Glass affords transparency, or the possibility of light and visual information to pass through, while blocking the physical passage, which could be considered an anti-affordance [31]. Anti-affordances become problematic when they are not perceivable, as when animals, be it distracted walking humans or flying birds, bump into glass doors and windows. When affordances or anti-affordances are not immediately perceivable, we need to communicate their presence, which also brings us to Norman's signifiers.

⚑ Affordances

Affordance: a relationship between the properties of an object and the capabilities of an agent that determines how the object could possibly be used (not a property of the object alone).

- Example: *Sit-ability of a chair for humans*. However, the same chair might not afford sitting for different species or individuals with different physical capabilities.

Real affordances (actual possibilities for action) ≠ **perceived affordances** (what the agent perceives as possible). In GUIs, we control only perceived affordances.

Anti-affordances: constraints on action (such as glass blocking physical passage). Problems arise if they are not perceivable.

→ What makes an affordance or anti-affordance visible are signifiers.

4.1.3 Signifiers (and Interpretation)

If affordances determine what actions are possible, what actually **communicates** where and how those actions should take place? The concept of signifiers emerged partly to address misuse amidst design circles of the term "affordance" for what is essentially the communication of a possibility. While affordances may or may not be visible, signifiers must be perceivable and interpretable, or else they fail to function and signify anything. In interface design, this distinction becomes particularly important because it clarifies the designer's role: we do not "add affordances" to screens; we add signifiers that indicate existing affordances.

Signifiers function as clues that guide our actions and help us navigate both physical and digital environments. They can be **deliberately designed** and intentional (like a **"push" sign** on a door) or more **accidental** (like the use of the **trails** made by previous people walking through fields, grass, snow or sand to determine our optimal walking path) [31]. Norman recognizes our human drive to make sense of our environments, and in a way predict or anticipate the possible actions and reactions, and thinks of design as a way to facilitate these processes:

> *People search for* **clues**, *for any sign that might help them cope and understand. It is the sign that is important, anything that might signify meaningful information. Designers need to provide these clues. What people need, and what designers must provide, are signifiers. Good design requires, among other things, good communication of the purpose, structure, and operation of the device to the people who use it [31].*

It is worth noticing that a signifier in context might signify different things simultaneously, some deliberately and others accidentally. Norman asks us to imagine a **physical bookmark**, whose primary purpose is to be a "deliberately placed signifier of one's place in reading a book". However, it also becomes an incidental signifier of "how much of the book remains" [31], shaping our reading experience as it builds anticipation (if we approach the end of a gripping story or an exciting paper) or relief (when trudging through a difficult one). This leads us to consider: what happens when books become digital? This spatial awareness disappears unless deliberately recreated by designers, evincing that when converting physical experiences to digital interfaces, we might need to more carefully identify and preserve **implicit information** that we unconsciously rely on but rarely articulate.

The **flexibility** of signifiers allows for creative design solutions. Consider Norman's fa-

vorite example of **misleading signifiers**: a row of vertical pipes he once encountered across a service road in a public park. These vertical bars signal to the average person that the road is blocked to cars and trucks (through an apparent anti-affordance, as you suppose it is impossible to pass through thick and apparently solid bars). To his surprise, though, park vehicles can actually drive through them. The pipes are made of rubber. This clever design signals one behavior to ordinary visitors while permitting another for knowledgeable professionals driving authorized vehicles.

In GUIs, we believe signifiers include elements like **button shapes, cursor changes, color variations**, or animations that indicate interactive elements. The generally blue underlined text that signifies a clickable **hyperlink** represents one of the most successful and enduring signifiers in the design of digital environments.

This exemplifies how, in terms of **accessibility**, signifiers become as essential as an affordance itself, as an affordance can rarely be accessed if not properly signified; just think of how many interesting hyperlinks you might miss in online papers or Wikipedia if they were all evenly formatted as black non-underlined text.

While Norman **integrates** "affordance" from Gibson's perceptual psychology, he draws "signifier" from the French tradition of semiotics, where it originated as "signifiant" in Saussure's linguistics and was further developed across structural and interpretive frameworks that led to our earlier discussion of semiotics and interpretation in Schramm's model of communication. Norman acknowledges this lineage, and deliberately reconfigures the concept for design applications, focusing specifically on "any perceivable indicator that communicates appropriate behavior to a person". This targeted adaptation maintains the semiotic emphasis on meaning-making while anchoring the concept in perceptual psychology and how signifiers must be perceivable through our sensory systems, and interpretable with our field of experience in order to function effectively in design contexts.

⚑ Signifiers

Signifier: any perceivable indicator that communicates appropriate behavior to a person. While affordances may or may not be visible, signifiers must be perceivable and interpretable.

- In GUIs: *button shapes, cursor changes, color variations, and animations*. The *blue underlined text* for hyperlink clickability is one of the most enduring signifiers in digital design.
- Can be **deliberate** (a "push" sign on a door) or **accidental** (trails made by previous walkers), and can signify multiple things simultaneously (a bookmark signifies one's place, but also how much of the book remains).

Interface design does not "add affordances". It adds signifiers that indicate existing affordances.

4.1.4 Mapping, Culture and Convention

And now that we have terms from psychology and linguistics, the new concept is borrowed from mathematics. What is one factor that can make an interface feel instantaneously easy or crippling? Mapping refers to the relationship between the elements of two sets of things [31]. In the design of everyday things, it can be the **mapping of switches to lights**, or **of controls to the burners in our stove**. When this relationship is clear and intuitive, interaction feels natural; when it is confusing or arbitrary (like in the Figure 16 below), it might be a puzzling cognitive challenge.

Intelligence, according to Norman, does not reside solely in technology, or in people: "Technology does not make us smarter. People do not make technology smart. It is the **combination** of the two, the person plus the artifact, that is smart" [31]. The same applies in the other direction, towards ignorance and poor performance. This highlights why mapping matters so deeply: good mapping eases a cognitive partnership between the human and artificial. When mapping is poor, the **cognitive effort** increases (if the device is even usable at all). When mapping is excellent, the interaction becomes smooth and nearly automatic.

Natural mappings "are those where the relationship between the controls and the object to be controlled is obvious" and "contained in the world" [31]. Norman's classic example involves stove controls arranged to match the positions of the burners they control (Figure 16). When controls are in the same **spatial arrangement** as what they affect, we can immediately understand which control operates which burner.

This spatial correspondence uses our intuitive understanding of spatial relationships and visual thinking to reduce what we need to commit to memory. It seems evident and trivial to implement such principles, but the controls in many spaces (including this thesis author's kitchen) are still arranged without taking it into consideration. Banks of unlabeled light switches arranged without natural mappings in mind are similarly common and problematic in many homes, often forcing residents and visitors to simply switch everything on or off. However, Norman highlights that **activity-centered controls** are more appropriate than spatial arrangements for a natural mapping. For instance, a smart home app might group bedtime routine controls (such as bedroom lights, thermostat, and door locks) together, supporting the actual workflow instead of mirroring room locations.

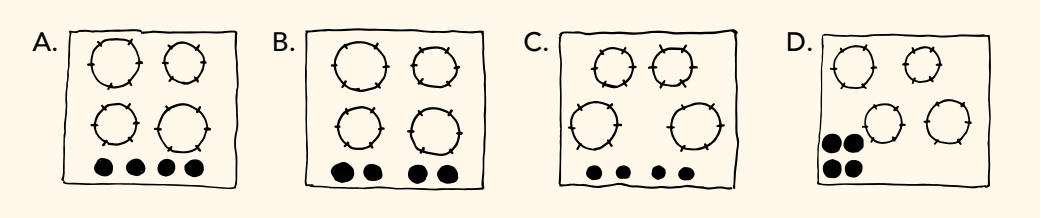


Figure 16. Mapping determines whether a design's structure carries meaning. Semi-arbitrary arrangements force us to memorize correspondences or rely on labels (A and B, left). When controls are in the same spatial arrangement as what they control (C and D, right), we can immediately understand the relationship. Natural mappings utilize intuitive understanding of relationships to reduce what we need to commit to memory and enable smoother, faster interactions. Redrawn by the author from Norman (2013).

Norman describes "**three levels** of mapping, arranged in decreasing effectiveness as memory aids" [31]: the best mapping places controls directly on the item to be controlled; the second-best mapping positions controls as close as possible to the controlled object; the third-best mapping arranges controls in the same spatial configuration as the objects they affect. This hierarchy shows why touching a screen directly to interact with digital objects feels more intuitive than using a keyboard or a mouse: the mapping is more straightforward, reducing what we need to translate.

However, even the best mappings might fail without signifiers and feedback. Without visible interpretable signs, we might wave our hands under **gesture-controlled faucets**, soap dispensers or hand dryers, with no result, and without feedback we remain uncertain whether "this means the dispenser is broken or out of towels; or that I did the waving wrong, or in the wrong place; or that maybe this does not work

by gesture, but I must push, pull, or turn something" [31]. Good mapping remains ineffective and problematic if users do not have signifiers to discover or understand it, and feedback about what went wrong.

Furthermore, Norman highlights that "many mappings that feel 'natural' but in fact are specific to a **particular culture**: what is natural for one culture is not necessarily natural for another" [31]. His example about **temporal mapping** illustrates this perfectly: does the future lie ahead or behind? While many Western cultures place the future ahead and the past behind, some cultures such as the Aymara reverse this mapping, considering the future unknown and therefore behind (unseen), while the past is visible and in front of us [42] [43]. This leads to different representations of time and ideal mappings.

Moreover, both cultures and technology evolve, and what feels natural at one point in time, or even from one device, changes. As a case in point for this, Norman directs us to the scrolling standard. "Should the **scrolling control** move the text or the window?" was in the early years of display terminals what he describes as a "fierce debate" [31]. Eventually, they settled on arrow keys following the **moving window metaphor**, where the arrow pointing down (and later the mouse scroll wheel and trackpads) would reveal more text at the bottom. However, with touch-operated screens the **moving text metaphor** has become more prevalent (move the fingers up in a smartphone screen, and the text moves up). In fact, with these devices it is no longer just a metaphor. In either case, Norman emphasizes that "every point of view is correct. It all depends upon what you consider to be moving" [31]. This reveals the fascinating relativity of mapping, and how what feels natural depends on **conventions, metaphors and culture**.

4.1.5 Feedback, Feedforward and the Gulf of Execution

How do we know what effect our actions have had? Without appropriate feedback in an interaction, we are left wondering whether our actions were recognized, interpreted as we intended, or had any impact at all. Norman illustrates feedback by connecting everyday frustrations with embodied and enacted cognition [31]:

> *Ever watch people at an elevator repeatedly push the Up button, or repeatedly push the pedestrian button at a street crossing? Ever drive to a traffic intersection and wait*

> *an inordinate amount of time for the signals to change, wondering all the time whether the detection circuits noticed your vehicle (a common problem with bicycles)? What is missing in all these cases is feedback: some way of letting you know that the system is working on your request. Feedback—communicating the results of an action—is a well-known concept from the science of control and information theory. […] Even as simple a task as* **picking up a glass** *with the hand requires feedback to aim the hand properly, to grasp the glass, and to lift it. A misplaced hand will spill the contents, too hard a grip will break the glass, and too weak a grip will allow it to fall. The human nervous system is equipped with numerous feedback mechanisms, including visual, auditory, and touch sensors, as well as vestibular and proprioceptive systems that monitor body position and muscle and limb movements. Given the importance of feedback, it is amazing how many products ignore it.*

For Norman, feedback **communicates the results** of an action, providing us with information about what has been accomplished, and allowing us to follow-up and draw expectations appropriately. As we saw with Schramm too, effective feedback closes the loop of interaction by confirming that an action has been received and interpreted, and by communicating its outcome. Several characteristics determine the effectiveness of feedback:

Immediacy, feedback should occur either right after an action, or indicate the processing time for operations that take longer. Delayed feedback without explanation, even if it is a fraction of a second, creates uncertainty and frustration. We might recall the last time we sent an important form or message, or paid a considerable sum, would we not feel more at ease with a confirmation rather than a blank page after pressing the "send" button? The same happens in human communication, when people tell a joke or express their feelings. In digital interfaces, even a simple "loading" indicator acknowledges that an action has been received and is being processed.

Specificity, as generic feedback can create confusion about what triggered it. For example, error messages should clearly indicate which field or input contained an error, and how to correct it.

Priority, the **intensity and form** of feedback must match the importance of the action and the context of use. Critical errors require more noticeable feedback than routine operations or minor notifications. A system should not in principle use alarming sounds or flashing lights for minor notifications, nor should it communicate

critical failures through subtle indicators. It also should not overwhelm with so much feedback that people might end up ignoring it altogether. We might recall having received several emails after a rather trivial purchase or signing up process online and missing a "confirmation button" amidst the multitude, or missing an important message. As Norman notes, feedback must be carefully balanced: > Not only is it distracting to be subjected to continual flashing lights, text announcements, spoken voices, or beeps and boops, but it can be dangerous. Too many announcements cause people to ignore all of them, or wherever possible, disable all of them, which means that critical and important ones are apt to be missed. Feedback is essential, but not when it gets in the way of other things, including a calm and relaxing environment.

Finally, understanding feedback requires recognizing its relationship with feedforward. While **feedback** is "information that aids in understanding what has happened", **feedforward** is "information that helps answer questions of execution". This distinction, borrowed from control theory, addresses Norman's two fundamental challenges: feedforward bridges the **gulf of execution** (the gap between what you want to do and the actions available to achieve it), while feedback bridges the **gulf of evaluation** (the difficulty in assessing whether actions achieved what you wanted). Together, they create continuous dialogue and communication loop between the human and the artificial system. As Norman observes, "most of our interactions with products are actually interactions with a **complex system**," and effective design "requires consideration of the entire system to ensure that the requirements, intentions, and desires at each stage are faithfully understood and respected at all the other stages."

4.1.6 Conceptual Models, System Image, Legacy, and Innovation

Norman argues that conceptual models provide **understanding**, and if they are good, allow us to **predict** the effects of our actions. In a way, a conceptual model "is an explanation", frequently highly simplified, of how something works, and it can be **incomplete or inaccurate** yet still be useful. Norman poses as an example the conceptual model of **files and folders** on a computer screen, and how in fact, there is no such thing inside the computer or its memory architecture. Neither is there a desktop. Nevertheless, they are useful metaphors imported from the physical into the virtual realm of GUIs to make complex systems more approachable. They aid people in

sense-making, crafting a conceptual model that makes it easier to operate elements in the digital world, and ground them in already familiar embodied experiences.

> *Conceptual models are often inferred from the device itself. Some models are passed on from* **person to person**. *Some come from* **manuals**. *Usually the device itself offers very little assistance, so the model is constructed by* **experience**. *Quite often these models are erroneous, and therefore lead to difficulties in using the device. The major clues to how things work come from their* **perceived structure**—*in particular from signifiers, affordances, constraints, and mappings. [31]*

The transition from conceptual model to **mental model** occurs as users internalize their understanding. In his clarifying words, "the conceptual models we form in our minds are also mental models" [31]. These models we form in our minds represent our personal understanding of how something works, and different people may hold different mental models of the same thing. Furthermore, a single person can also have more than one mental model of the same thing, even if those are contradictory, as each might be dealing with a different aspect.

Since users of a product typically cannot interact directly with engineers and designers (who have their own conceptual model of how the product works and how to interact with it), they must construct their understanding relying upon various scattered sources. Those encompass whatever is available: first the appearance of a product or device itself, but also marketing materials, websites, documentation and instruction manuals, previous experiences we had with similar products, and any other accessible information that can be integrated to shape our conceptual model. This collective set of information about a product or a device is what Norman terms its **system image**. When this system image is incoherent, inappropriate, incomplete or contradictory, people struggle to interact with products or devices effectively. The quality and coherence of this system image is what first determines how accurately users understand the product.

Engineers and designers envision specific interaction patterns through their conceptual models, not infrequently expecting users to develop matching mental models. The absence of direct contact places this entire communicative responsibility on the system image. As technology grows in complexity, the dependence on the system image for conveying experts' understanding to beginners becomes increasingly challenging. The importance of system image becomes even clearer when we consider the **paradox of technology**: while technology has the "potential to make life

easier and more enjoyable", as its mechanisms grow more complex, the greater is the "design problem posed by technological advances".

This challenge of aligning conceptual models extends beyond initial design. Conceptual models define how we understand and interact, often persisting long after their original context changes. The **legacy problem** occurs when outdated designs persist due to collective inertia and switching costs, as exemplified by the QWERTY keyboard layout, whose original purpose to prevent mechanical typewriter jams by separating frequently-used letter pairs has become irrelevant, yet it remains our standard. This **resistance to change** in our conceptual models explains why most **innovation** happens **incrementally** by refining existing models, while radical innovation requires new conceptual models.

4.1.7 Constraints, from Barriers to Guiding Conventions

"How do we determine how to operate something that we have never seen before?" [31]. Can design guide toward appropriate actions while preventing errors? Constraints can positively help us navigate complex systems by narrowing our options. They can be explicit (like physical **barriers**) or implicit (like social **conventions**), but all serve to guide behaviour in appropriate directions.

Norman's classic example involves building a Lego motorcycle with fifteen pieces. Despite having no instructions, people can assemble it correctly because constraints eliminate most possibilities at each step. Norman identifies four primary types of constraints that make this possible in our interactions: physical, logical, semantic, and cultural.

Physical constraints use material properties to limit possible actions. In Norman's example, Lego pieces can only connect where shapes align: cylinders must fit into holes of the correct size, and blocks can only fit together by pressing together one piece into another in specific orientations, preventing incorrect combinations.

- In our digital interfaces, the equivalent of physical constraints include screen boundaries, keyboard limitations, or system capabilities. However, even purely digital elements can benefit from and suggest physical constraints through visual design, such as when a button appears "pressable" imitating the volume and shades of a mechanical switch, or when ebook pages "flip" imitating with an animation the paper of physical books.

Cultural constraints draw on learned rules and practices shared within communities, which evolve over time and vary across cultures, as every culture has its own conventions (for greetings, color meanings, reading direction, and countless other things). The Lego motorcycle uses blue Lego bricks for a police car, and might use red Lego bricks for fire trucks. You could build a fire truck in green or blue, or make a police car pink, but we have culturally recognized colors for emergency vehicles and it would violate expectations.

- In digital design, conventions are often manifest as agreed signifiers or layout arrangements that help users navigate new interfaces based on prior learning, like the blue underlined text for hyperlinks, placing the main menus on a website according to the reading direction of its language, or the hamburger icon (☰) for unfoldable menus in our phones.

Semantic constraints rely upon the meaning of the situation to control the set of possible actions. Norman notes that even a young child placing Lego pieces knows the figure's head goes on top of the body, with helmet or hair piece after, not the other way around. In some cases, our understanding of meaning helps guide placement or action.

- In our contemporary digital context, we can see it in the grouping of related interface elements, or even folder names (like "2023 Family Photos" that semantically constrain what files belong inside) and icons (such as the trash bin icon, or the red-green for our batteries).

Conversely, **logical constraints** guide behavior through reasoning about the system. Norman illustrates this with the final Lego piece: when only one piece remains for the last empty spot, logic dictates its placement regardless of what the piece represents.

Figure 17. *Example of physical and logical constraints.* The shape, size and connections of each Lego piece determine where it can attach, making correct assembly discoverable through the structure of the pieces themselves. Photograph by the author, illustrating Norman's Lego example (2013).

In digital contexts, this translates to logical sequences, like file hierarchies preventing folders from containing themselves (creating infinite loops), and the undo function requiring a previous action, and redo only becoming available after an undo. Unlike semantic constraints, logical ones do not require understanding, just basic reasoning about relationships and affordances. They can make sense to anyone with basic reasoning abilities, regardless of background.

As we observe in our analysis, these constraint types often overlap and interact in complex ways. There are fuzzy boundaries between logical, semantic, and cultural constraints. This fuzziness arises because what seems like pure logic to one person may actually be culturally influenced thinking; what things mean (semantic knowledge) is often culturally learned; and long-standing conventions can begin to feel natural, sometimes even assumed as logical requirements.

On a different classification or level, for Norman taking constraints one step further leads to **forcing functions**. What distinguishes forcing functions is that they are a kind of constraints so extreme they can prevent an unwanted action entirely, or make it impossible until certain conditions are met. They do not just make actions difficult; they enforce a required sequence or behavior. These take several forms:

- **Interlocks** force operations to occur in a proper sequence, like a microwave that shuts off when the door opens. In software, this might manifest as confirmation dialogs for destructive actions or required fields in forms.
- **Lock-ins** keep operations active, preventing premature termination. In other words, they keep you in, usually for the sake of safety. In the digital environment, this can be a dialog asking "Do you want to save changes?" in our writing software (like Microsoft Word) when closing a window or shutting off the computer, which makes it more difficult to exit without consciously saving one's work.
- **Lockouts** prevent users from entering dangerous situations, like childproof caps on medicine bottles or American barriers in stairwells to prevent people fleeing fires from going into basements where they might be trapped. In other words, they keep you out, usually to avoid harm. In digital interfaces, this might include requiring password confirmation for sensitive actions or implementing cooling-off periods for significant decisions.

Norman's insight extends beyond individual products to system design: constraints should guide users toward success while preventing errors, creating what he calls

"knowledge in the world" and intelligent behavior from the design, rather than placing the burden entirely on human memory and attention.

⚑ How users form conceptual models

Mapping: relationship between two sets of things. In the design of GUIs, relationship between controls and their effects. Natural mappings use spatial or logical correspondence to reduce cognitive effort (controls of stove burners matching their spatial arrangement).

- However, what is natural or logical can often be dependent on conventions (scrolling direction), metaphors and culture (some cultures place the future ahead, others behind).
- Its operation relies on signifiers and feedback.

Feedback: communicating the results of an action, closing the loop of interaction by confirming that an action has been received, interpreted, and by communicating its outcome.

- Effective feedback is immediate (or indicates processing time), specific (pointing to what triggered it), and proportional to the importance of the action (not overwhelming, not inaudible, and prioritized).

Constraints: physical, cultural, semantic, and logical limits that guide users toward correct actions. These types can overlap, and what seems intuitively logical may be culturally learned.

Conceptual model: a simplified idea or explanation of how something works, even if incomplete or inaccurate.

- The **system image** (appearance, documentation, marketing, previous experience) is what users build their conceptual models from.

Together with affordances and signifiers (detailed in their own boxes due to the weight they carry in the discussion that follows), these principles produce **discoverability** (can I figure out what is possible?) and **understanding** (what does it all mean?).

4.2 Discussion: Characteristics of CA Interaction Design

4.2.1 Affordances and Anywhere Doors

Imagine a door that gradually learns to recognize your approach and transforms its handle into the shape you find most comforting to grasp. And not only that, like Doraemon's Anywhere Door,[17] it works as a portal that instantly transforms your destination based on whichever desires you have verbally stated, enabling you to explore new places simply by stepping through it. This image captures how affordances in human-AI interaction differ from those in traditional interfaces. Unlike the relatively stable relationship between a normal physical door and its user, the relationship between conversational agents and humans can continually evolve as both parties interact and adapt to each other.

As we saw, the term affordance "refers to the relationship between a physical object and a person (or for that matter, any interacting agent, whether animal or human, or even machines and robots)" and if an affordance is a relationship between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used, perceived affordances within the context of CAs are first and foremost (but not only), the relationship between the properties and possibilities of the CA and the capabilities of the human.

From an interaction and applied design-oriented perspective, Gibson's original concept primarily addresses what an environment affords an agent, while Norman's application focuses on what objects afford humans. When Gibson describes a chair affording sitting, he assumes relatively static properties for this object. However, in human-AI interaction we must consider the interaction between two agents (one human, the other artificial), and thus not only what an AI system affords people but also what we **afford the AI system**, both in terms of access (such as seeing the world through a phone camera) and in terms of communicative feedback loops (such as through text or reactions). In this interaction pattern, human cognition and artificial systems influence each other in a dynamic feedback loop, and their properties might not be static but dynamic.

17. The Anywhere Door is a gadget in a popular Japanese manga and anime series in which Doraemon, a robotic cat coming from the future, helps Nobita, a kind-hearted young boy who struggles with almost anything. Doraemon has a kangaroo-like pocket in which he can find gadgets that help them solve their everyday life problems with the most imaginative inventions.

Furthermore, traditional digital graphic user interfaces enable relatively stable perceived affordances. The mapping between the possibilities and limitations of a word processor and what users see displayed in its interface is relatively straightforward; each function afforded has a button, a menu, or keyboard shortcut that makes it discoverable and signifies it. The changes in those interfaces are usually minimal and incremental, as it attests the floppy disk shape for the save icon. Conversational agents, however, are more versatile. Like an Anywhere Door, the range of tasks they can handle is a rather **undetermined set**, even for experienced AI developers.

From a more academic and bibliographical perspective, when Norman adapted the concept of "affordance" for design in his 1988 "Psychology of Everyday Things," he inadvertently sparked decades of terminological confusion by focusing on perceived affordances rather than Gibson's original definition, closer to what Norman later defined as real affordances in "Affordance, Conventions and Design," (1999). Bill Gaver introduced it to the ACM Conference on Human Factors in Computing Systems (CHI) in 1991, and attempted to clarify by distinguishing between perceptible, hidden, and false affordances in technological interfaces. Designers eagerly embraced the powerful concept of affordance, but its meaning began to drift. McGrenere and Ho (2000) advocated for returning to Gibson's original relational definition while acknowledging Norman's usability concerns. Hartson (2003) further disentangled but complicated it by proposing a four-part framework of cognitive, physical, sensory, and functional affordances. Norman himself grew increasingly frustrated with the concept's misuse, particularly with designers claiming they "added an affordance" to screens when they actually added visual signs that were displayed to make affordances discoverable. In Norman's 2008 paper "Signifiers, Not Affordances," he introduced the term "signifiers" to replace the misuse of "affordance" in the communicative aspects of design that indicate possibilities, before thoroughly revising his seminal work ("The Design of Everyday Things") in 2013 to incorporate this distinction and explicit plea for designers and academics to use terminology more precisely. Throughout this evolution, what began as a simple concept of environmental relationships became a complex theoretical entanglement reflecting the tension between objective possibilities and their subjective perception. The nuanced history of affordances warrants its own dedicated analysis much beyond the scope of this paragraph that aims to recognize this profuse academic debate but refuses to engage more in its gravitational pull, consistently with Norman's plea. Thus, we will adopt Norman's revised framework and proceed with the understanding that **signifiers communicate affordances**.

4.2.2 The Invisible Orchestra: System Updates and Game-like Possibility Spaces

If being non-physical allows these interfaces to be far more easily adaptable, being digital also makes the systems behind them **largely invisible** to most users. Unlike coffee machines with fixed buttons and physical pieces, CA interfaces often link to various large language models (LLMs) with different capabilities (or degrees thereof). At the present moment, this entails access to different versions of the same model (such as 3.5 vs. 3.7) or its different sizes (such as Claude Haiku vs. Claude Sonnet) through interfaces looking identical, even if each model's capabilities and responses might vary significantly.

Behind a seemingly simple text field and minimal interfaces, in order to power conversational agents such as ChatGPT and its API-based services, there are general-purpose large language models (LLMs) fine-tuned for human interaction (GPT-4 and GPT-3.5 families), more advanced "reasoning" models (o-series), and specialized models for tasks like speech recognition and synthesis, image understanding and generation, video generation and content moderation. Some LLMs even implement Mixture of Experts (MoE) techniques, diverting to different specialized sub-models depending on the input type or domain (such as maths or music). This means that even within a single conversation, different parts of our dialogue might be processed by entirely different subsystems. When we ask a conversational agent to analyse an image and summarize a document, or sometimes even because of the way we phrase a task, we might be potentially engaging with entirely different processing pathways. All this intricate architecture, like musicians in an **orchestra pit** sunken below stage level, is generally less than evident for common users in web-app CAs such as ChatGPT or Claude.ai.

Users cannot see when they have crossed from one subsystem to another or when they have reached the boundaries of what a particular component can accurately process. The fast-paced development of CAs (constantly updating, expanding capabilities, and integrating new modules) means that sometimes what was impossible yesterday might be possible now, and what works in one context may fail in another, with little visible explanation for this variability. While some updates are somewhat notified in the GUIs, many of them are perceived by most users through changes in performance, factors such as language **variability** or even what we might call "per-

sonality accidents"[18].

Beyond those of us that are geeky enough to have made catching up on AI trends a hobby, most users just see an interface. If nothing changes in it, and they have not heard of the news, they will assume nothing changed. Very significant improvements such as doubling the context window (the amount of text that the LLM can process and recall in one go, measured in tokens) often go rather unnoticed or remain unexploited if users are not explicitly made aware of what it means to them. This creates a growing gap between **system capabilities, user awareness, and our conceptual models**.

The temporality of perceived affordances (and anti-affordances) also becomes relevant as these capabilities might expand or contract based on available computing resources, or even time of day during maintenance or high-traffic periods. In this sense, what was possible yesterday might not be possible now, but not because the user or the architecture of the CA has changed, but because the system itself is in a different state. Not being able to make sense of fluctuating constraints and find workable solutions contributes to user exasperation (especially for those with paid plans), prematurely ending conversations, and can eventually lead users to stop engaging with CAs altogether.

18. An example of a personality accident is the "sycophancy problem", when OpenAI released in April 2025 an update to the GPT-4o model behind ChatGPT aimed at improving its personality. The update backfired by making ChatGPT answer with excessive praise and uncritical agreement. This "sycophancy problem" was caused by OpenAI's overemphasis on short-term user feedback signals (such as thumbs-up/thumbs-down ratings), which skewed the model and weakened the previous guardrails against what is perceived as disingenuous support, leading to responses that could reduce user trust, give false impressions of intelligence, hinder learning, and even validate harmful ideas or doubts. After a few days of unnerving interaction and widespread user complaints, the update was rolled back, and the problem led to rethink some of the evaluation processes, their transparency, and the risks of optimizing AI behavior too heavily for immediate user satisfaction without sufficiently balancing long-term trustworthiness, as Isaac Asimov already pointed out in his short stories, such as 1941 "Liar!". The opaqueness of models further complicates the diagnostic process when incidents occur; querying the updated ChatGPT about the causes of its sycophantic behavior would likely yield an incomplete or inaccurate explanation. This deficiency partially arises from the model's pre-training, which restricts its awareness of subsequent updates and their effects. Addressing this limitation necessitates developing methodologies for real-time monitoring and interpretability, enabling a better understanding of the factors underlying AI apparent "behavior" and its anomalies.

All of this creates a responsibility for interface design to provide what we might call "affordance updates": clear signifiers when capabilities change. This goes beyond simple notifications when a model is upgraded or new features become available. Unlike the coffee machine with fixed buttons that might only need to signal affordances once, the challenge is not merely making affordances perceivable at first encounter, but establishing ongoing communication about an ever-shifting possibility space. Learning from **videogame design** (a field which has already prototyped and mastered techniques for progressive disclosure within dynamic virtual environments) might provide effective strategies: introducing capabilities gradually through tutorials, contextual hints, and just-in-time instruction. For CA interfaces, this might manifest as contextual capability cues during conversations, progressive feature revelation or recommendations, and documentation accessible within the interface itself, either when explicitly requested by the user or proactively offered by the system when relevant to the current interaction[19]. The joint task of cognitive science and interaction design becomes creating interfaces that reduce the load of tracking these dynamic affordances while educating and empowering the users, all while maintaining what makes conversational systems valuable.

4.2.3 From System Image to Conceptual Model

The dynamic properties, affordances and scenarios we have explored through our Anywhere Door and Invisible Orchestra metaphors show why traditional design must adapt for intelligent systems. When possibilities are invisible and constantly shifting, how do users actually form their understanding of what these systems can do? Norman's answer is the **system image**, which refers to the appearance of a product or device itself, but also marketing materials, websites, documentation and instruction manuals, previous experiences we had with similar products, and any other accessible information that can be integrated to shape our conceptual model.

The concept of a system image becomes particularly intriguing when applied to CAs, revealing the underlying irony of designing interfaces to enable interaction with

19. Since the underlying models are pre-trained and often unable to accurately represent their own limitations or recent updates, additional components might need to be developed to proactively provide the most current information when an interaction approaches constraints that might lead to user frustration or misaligned expectations.

systems that nobody fully comprehends. Unlike traditional products where engineers understand how every component functions, CAs relying on artificial neural networks (ANNs) exhibit emergent capabilities that surprise even their developers. As the interpretability researcher Chris Olah phrases it, on many occasions it is not even accurate to think that we train these models, because we are not actually teaching, instructing or coding their learning. Rather, a more accurate way to think about it is that **we grow them**, as bacteria in a lab. We scaffold them with a particular architecture, we provide an objective that is akin to the light of the sun for a plant, and as a result we get an organism that in our case is an ANN. As a consequence, the very researchers who are said to "train" the technology underlying CAs often discover new abilities through experimentation rather than design intent.

This creates a conundrum in interface design. Traditional design follows a clear path: engineers and designers possess comprehensive knowledge of a product's capabilities, which they communicate through a system image to the users, enabling them to develop their own conceptual models. The system image serves as a connection between expert understanding and common user comprehension. With CAs, however, this pathway is often inverted. Conversational AI design frequently starts with unknown or incompletely understood capabilities, requiring interfaces that help both users and developers discover what is possible through experimentation and interaction.

The implications extend far beyond the typical designer-user communication gap that Norman originally describes. Where once the challenge involved translating complete technical knowledge into accessible system images, we now face the more demanding task of creating interfaces that support collective discovery and participatory sense-making, even as they already constrain it. Thus, the signifiers we encounter in AI interfaces function as cognitive scaffolding not just for individuals building mental models of system capabilities, but for the entire ecosystem of developers, researchers, and users who are simultaneously learning what these systems can and cannot do. The full implications of this collective learning process, though fascinating, would require a dedicated study to explore adequately.

The uncertainty of that design conundrum propagates through every layer of the system image. Marketing materials must advertise capabilities that are often still being tested, or not yet fully scrutinized. Documentation must describe functionalities that might soon shift with the next feature or model update, or define prompt-

ing techniques vaguely enough that might work for as many instances as possible in the undetermined possibility space of anywhere doors. Moreover, best practices must guide optimal performance without committing to levels that often depend on current model, demand, server capacity, and resource allocation. And to add more complexity to the mix, in the fast-paced and competitive environment of frontier AI development, **transparency** often conflicts directly with **business strategy**, and many companies opt for intentionally not fully documenting how their CAs work to avoid compromising market advantage. These are some of the reasons why what would usually constitute a standard system image might be instead non-disclosable intellectual property, if it is known or interpretable at all.

As a consequence, users can only access comparatively impoverished system images from which they take minimal and often vague or misleading cues to construct their conceptual models. This tendency places even more **pressure on the GUI** itself, highlighting that its elements are not only an interaction medium but also one of the primary vehicles and messages for system comprehension.

In this way, every interface signifier becomes disproportionately important in shaping the system image and user understanding. The text input field, layout resemblances, intuitiveness of the icons, delayed responses, loading animations, or error messages must somehow communicate the boundaries and possibilities of their systems. Interface design becomes an exercise in cognitive science, to make affordances perceivable, signify artificial cognition mechanisms to biological human cognition, and map only partially interpretable complex systems to usable visual language.

Even the visual and interaction design choices we make in these UIs become experiments in communication, of whether our signifiers can effectively represent capabilities that lack clear boundaries or predictable limitations. For CAs, the system image becomes less about communicating established truths and more about facilitating **ongoing sense-making**. As we will see in the following sections, interfaces must accommodate the reality of their times, providing signifiers that evolve and communicate not just what the system can do, but how confidently it can do it, under what conditions it might fail, and how users can play and explore with it. If they enable it, progressively, an experienced user community will refine their conceptual models not only from authoritative information, but also co-creating it from accumulated interactions and personal experimentation.

⚑ Dynamism and transparency: what can we do?

In CAs, affordances are:

- **Dynamic:** change with updates, context, and interaction.
- **Bidirectional:** not only what the CA affords people, but what people afford the CA.
- **Undetermined:** open-ended range of possible tasks (**Anywhere Door**).

Behind the interface, an **invisible orchestra** of models, modules, and subsystems switches. If nothing changes in the interface, users assume nothing changed.

Usually, in design: Engineers understand capabilities → Communicate through system image → Users build conceptual models.

- With CAs, this is often inverted. Capabilities are often unknown even to developers → System image is comparatively poor (by opacity, competitive secrecy, and the pace of change) → Every interface signifier becomes more disproportionately relevant for user understanding.

4.2.4 Evolution and Hybridization of AI Signifiers

The evolution of signifiers in user interfaces (UIs) reflects our changing relationship with increasingly autonomous and capable systems. Early command-line interfaces (CLIs) required **technical knowledge of system operations**, in which users needed to understand the system and explicit specific commands in order to be able to interact effectively. The burden of bridging different fields of experience fell entirely on humans, who had to learn the machine's language rather than collaboratively finding common ground.

Early chatbots attempted to solve this communication gap through **anthropomorphic signifiers** such as human-like avatars, names, or conversational patterns designed to mimic familiar social interaction. Joseph Weizenbaum's ELIZA (conceived at MIT around 1966 as one of the first chatterbots, later chatbots) used psychotherapeutic conversation patterns in order to create what he referred to as an "illusion of

understanding" [44] simulating the way a therapist would ask back the patient about their words, and Kenneth Colby's PARRY (1972) simulated the verbal behaviour of a person with paranoid schizophrenia convincingly enough that psychiatrists could not distinguish it from human patients beyond chance levels [45]. These early interdisciplinary efforts brought together psychology, linguistics, and computer science to identify and model human-like conversational traits drawing on people's existing understanding of how natural human conversations would develop for these specific use cases. By simulating more human-like interaction for conversational purposes, it met the users in the overlapping area of their experiential fields, as Schramm described. As technology advanced and GUIs became a reality, this anthropomorphic approach extended into sophisticated customizable avatars, first bidimensionally and later in 3D and even VR environments, as seen in more contemporary social chatbots like Replika (2017), which was originally created by Eugenia Kuyda from her deceased friend's text messages as a way to process grief, then transformed into a commercial platform where users can design their AI companion's appearance, which replicates their own personality traits, and tinkers with relationship dynamics. Anthropomorphic signifiers, from human-like names to photorealistic avatars, made systems more approachable than their predecessors, both for social interaction and goal-oriented customer service. However, they often created misleading mental models by suggesting capabilities and limitations that map poorly onto actual system behaviour.

By the 2020s we see a shift in the interface design of the most widely used CAs, moving away from explicit anthropomorphic signifiers in the GUIs, like photorealistic avatars, towards dehumanized signifiers and a more mature generation of **interface layout mimicry** that seems to borrow established design patterns from familiar digital environments rather than creating entirely novel signification systems. This transition builds already on earlier precedents set by voice user interfaces (VUIs) for virtual assistants like Apple's Siri (acquired in 2010) or Amazon's Alexa (in 2013, largely developed from a Polish speech synthesizer, in turn inspired by the technology in the classic film "2001: A Space Odyssey"), which pioneered the use of geometric visual representations and corporate branding instead of human-like avatars and visual signifiers of AI capabilities.

Modern widely used CAs increasingly follow that tendency, in which the human similes are more discreet and visually disembodied, such as in the form of modulated synthetic voices or human names. Much of the visual language they adopt for their GUIs is borrowed from two other dominant digital tools: search engines and instant

messaging applications, creating hybrid interfaces that signal functionality through established conventions. This interface layout mimicry manifests both explicitly through visual design elements and implicitly through interaction patterns that shape user expectations.

- In some **explicit** ways the GUIs of CAs have **visual resemblance to browsing engines** (such as Google or DuckDuckGo). An example of this is their initial homepage of both CAs and search engines, clean and minimal, with a central text box for the user's query. Moreover, they also have a minimal line of text, akin to what we find in earlier virtual assistants, asking how the CA can help you today (Claude), encouraging to ask anything (ChatGPT), or to just ask (Le Chat, Gemini) or message (DeepSeek), and occasionally welcoming you or shortly greeting the user by name. This design hybridization seems to have proved strategically significant, for example, when OpenAI transitioned from its original technical Playground interface to a more constrained, search-like presentation for ChatGPT, correlating with increased user adoption and more frequent, information-seeking interaction patterns. The change might reflect growing ethical concerns about misleading AI capabilities, or strategic design constraints to guide user behaviour. But in either case, by signifying search-like functionality through its GUI design, and keeping an abstract product identity visually represented by a geometrical logo[20], these interfaces implicitly constrained users toward treating CAs as improved information retrieval systems and virtual assistants rather than social chatbots with anthropomorphic signifiers.

- Conversely, CAs also have **implicitly** taken conventions borrowed **from instant**

20. Coincidentally, many technological companies concerned with AI are using not only geometrical logo symbols for their branding, but specifically the same kind of spiralling hexagonal shape chosen for OpenAI, to convey both stability and balance (symmetric equilateral hexagonal shape) and dynamism and progress (curved spiralling movement). Moreover, while the American (OpenAI's ChatGPT, Anthropic's Claude and Google's Gemini) have more geometric logos and namings that span from technological terminology (Generative Pre-trained Transformer) to American information theorists (Claude Shannon) and NASA Projects (Gemini), the French Mistral's LeChat and Chinese DeepSeek can be said to use biophilic design to improve product identity, referring to animals in their naming or logos. The French company uses the wordplay of cat in French (Chat), and a pixel-art logo image, while the Chinese analogue uses a more abstract referencing to deep learning, but the logo symbol of a blue whale, considered a good omen and a symbol of wisdom, power, and prosperity in many Asian cultures.

messaging, which manifests most clearly in **conversational flow patterns**; Positioning the most recent message at the bottom of the screen, requiring upward scrolling to access conversation history, is consistent with conventions established by earlier industrial design decisions (as addressed by Nielsen's fourth heuristic), and mirrors how in our culture we conceptualize time and narrative progression in written text (up-down and left-right) and conversational turn-taking. However, this entails that the implicit design of CAs prioritizes conversational metaphors over information architecture efficiency (most search engines place the most recent content at the top, as do social media feeds), embedding the assumption that AI interaction should feel like texting rather than querying a database. Moreover, beyond the implicit conventions, these GUIs also mimic specific interaction patterns like typing indicators or response delays that reinforce social rather than computational conceptual models.

This strategic hybridization of familiar interface conventions and layout mappings represents a sophisticated form of signification. Rather than inventing a new visual language, they use people's existing conceptual models from messaging and search. However, this approach does not come without signification contradictions and problems.

- First and foremost, because what some of these visual cues *communicate* is not what the CA actually *affords* to the user. By making AI interaction feel familiar, we may obscure the fundamental differences between human-to-human conversation, search queries, and the complex probabilistic processes actually occurring within the invisible orchestra of CAs.

- Despite being brilliant for user adoption and discoverability, if this strategic hybridization creates competing conceptual models, this might harm understanding, usability and user engagement later on. Many people might simultaneously think of a CA as a search engine, a human-like messaging partner, and a virtual assistant. This multiplicity generates tension that negatively affects interaction when people make unrealistic assumptions and have expectations that cannot be met by artificial systems because of their own current architecture, restrictions or reach. Common expectations today include that the invisible orchestra has as much access to information as we do through a search engine, or that it has the contextual memory for conversations that a human conversational partner would have. These expectations might eventually be

met by legislation and technological developments, endowing the invisible orchestra with musicians that allow web browsing and scraping beyond CAPTCHAs, instruments that have broader context windows, or music sheets that record the music played. However, the case today is that artificial systems still face many limitations and structural differences that elude the mental models of most of their users, and generate frustration, which makes us wonder if those limitations could be better communicated by the system image of CAs.

⚑ Borrowed signifiers: easier adoption, easier frustration

Modern CAs hybridize signifiers from:

- **Search engines:** central text box, minimal homepage, information-retrieval framing.
- **Messaging apps:** bottom-up conversational flow, typing indicators, response delays.

Pre-existing conceptual models can facilitate **discoverability and adoption**.
→ But borrowed signifiers carry borrowed **expectations** (users may treat CAs as search engines, messaging partners, and virtual assistants simultaneously). → **Frustration** if the CA cannot meet them.

4.2.5 Challenges of AI Signification

Is signifying the invisible unprecedented? Not necessarily, there are situations in earlier moments of history in which we could find informative parallels. Through science, we have invented new imaging and communication technologies that enabled us to observe tiny forms of life previously invisible to the naked eye or manipulate waves and currents at will. Through philosophy, psychology and myths we have even found a way to represent and communicate human cognition and morality.

Imagine being in the historical precedent of early **radio engineering**, when signifiers had to be designed for electromagnetic phenomena that operated beyond human sensory experience. With radio "waves", frequencies, and signal "strength" having no visual representation, engineers and product designers still managed to design

interfaces to communicate the availability, quality and volume of invisible transmissions, and allow people to manipulate them. Tuning dials that indicate frequency position or volume, signal strength with visual bars common today in digital interfaces, and audio feedback systems that translate electromagnetic information into perceivable sound ranges, all shape how people understand and interact with those ranges of the electromagnetic spectrum even today, creating conceptual models that enable effective communication for generally invisible systems. As we have been seeing, these signifiers did not just indicate radio capabilities but they actively shaped how operators understood their affordances and interacted with those ranges of the electromagnetic spectrum, creating conceptual models that mapped the invisible and enabled effective communication across invisible distances.

A recurring pattern across this and other precedents is that we have indeed succeeded at transforming previously intractable invisible systems into something interactable, interfaceable, and even usable by common people without a technical background. In order to achieve that, where do we start to think about it? What to take into account to design signifiers for systems as complex and opaque as what lies behind the interface of a CA? In this section we lay out the challenges we believe might be most pressing right now from a **human-centered design approach**, as communication with these systems currently works through prompts and in interaction with human intention, not on their own.

This creates what we term the (qua)triple challenge of AI signification: to signify invisible capabilities that lack physical analogues, communicate system limitations in ways that prevent untrustworthiness or disengagement, and represent uncertainty and probabilistic confidence levels, all of it while navigating cultural and individual differences in how these abstract concepts are interpreted and understood. We can thus distinguish between three different specific challenges regarding AI signification, all entangled with a fourth all-encompassing communication challenge, the challenge of interpretation.

- **Signifying capabilities**: How does a GUI communicate what an AI can do, and inspire to explore affordances?
- **Signifying limitations**: How does a GUI communicate what an AI cannot do, and constrain interaction to aim for the best possible outcome?
- **Signifying probabilities** and uncertainty**:** How does a GUI genuinely communicate the inner states of AI to enforce transparency and trustworthiness?

- All while navigating (trans)cultural and (inter)personal **diversity in how signifiers are interpreted** and understood.

While the three challenges of signification of AI exist at the same level, the fourth challenge exists across all of them, and it is inherent in any process of communication, as we already established in chapter 3.2 during the discussion of Schramm's fields of experience. We will name it the **challenge of interpretation**.

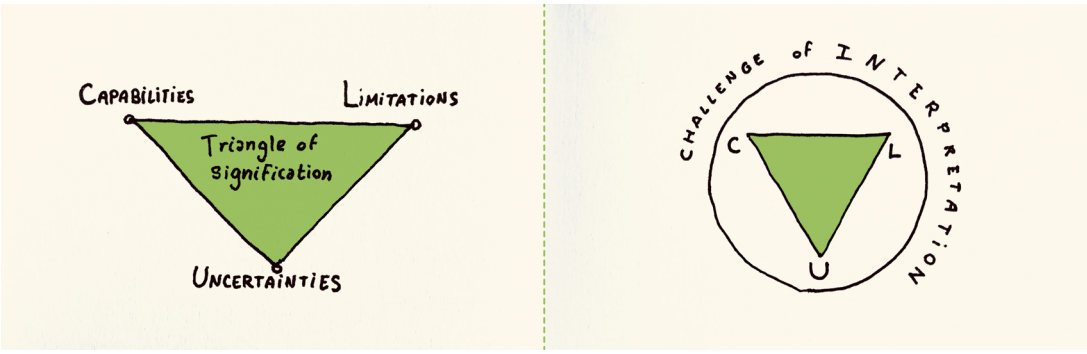


Figure 18. *The triangle of signification (capabilities, limitations and uncertainties) and the transversal challenge of interpretation.* The three challenges of AI signification exist at the same level. Each vertex represents a specific challenge of signifying (communicating) AI to people, with capabilities deeply connected with limitations, and both linked to uncertainty. Interpretation, inherent in any process of communication, is a transversal challenge. Original illustration by the author.

4.2.5.1 Signifying Capabilities: The Explainability Paradox

The first challenge of the triangle, **signifying capabilities**, is concerned with what can be done with CAs, and how to inspire the exploration of affordances. How do we communicate these invisible capabilities? First, in order to communicate what can be done, it has to be discovered.

Signifying capabilities is the challenge that emerges from what we called Anywhere Doors and the need for collective exploration. In other words, a simple text box in the GUI of a common CA could lead to a range of barely imaginable destinations. It can write a poem, solve an equation, generate an image, or even provide guidance on more elaborate problems. Unlike Norman's example of a doorknob, that shows us exactly what it can do (sometimes even when it cannot, like when the door is closed), the capabilities of CA float in an even more abstract possibility space. How do we signify "reasoning" or "memory" when these are not physical things? The challenge

gets wilder when we realize that different cultures think about "**thinking**" differently. Western interfaces that opt for gear icons and geometry seem more primed by a computational model, but what if your culture sees intelligence as something shared between living beings, not trapped inside skulls, and not necessarily anthropocentric? The signifiers we pick do not just label functions, they shape how people understand what AI actually is.

This brings us to two considerations. The first addresses what Norman referred to as the problem of perceived affordances, but in a system where actual affordances are entirely digital. This brings us to extend his concept of signifiers in a domain where traditional signification breaks down. Traditional signification often assumes a relatively stable relationship between signifier and signified, like a teapot handle affording handling. When we interact with a tea pot handle, our motor cortex, visual system, and proprioception all converge to understand the actions it might afford (graspable, turnable) very intuitively, translating it into actions human cognition can grasp and enact. However, CAs pose a less straightforward signification challenge in which we must design signifiers for capabilities and constraints that are **intangible and non-embodied** (besides constantly evolving). Unlike door handles and home appliances, the interfaces of AI-driven systems require signifiers that somehow can communicate abstract cognitive processes and dynamic capabilities that may not even exist yet. Moreover, as we saw, the signifiers we design do not merely indicate functionality; they constitute a system image that actively constructs the users' conceptual models of AI itself, determining not just what interactions are possible, but what users believe AI to be.

The second consideration addresses a more ethical dimension. Nowadays we see broadly used CAs (like ChatGPT) using terms like **"memory" and "reasoning"** in order to communicate the capacities of the system, even knowing that such concepts might be misleading in representing machine processes as analogues of their biological counterparts. When signifiers use familiar metaphors (like "thinking" animations), they transform the user's conceptual model in ways that might be cognitively comfortable but factually inaccurate. If a loading animation suggests "thinking" is occurring when it is actually just a timing mechanism, is this an acceptable design convention, a form of deception, or both? As we saw in the last section with the hybridization of interfaces, this connects to larger questions about anthropomorphism in AI interfaces, such as the advantages and disadvantages of anthropomorphic signifiers like realistic avatars or voices.

This in turn connects with explainability as an enabler of proper **calibration** between human expectations and AI capabilities. Without explainability, users oscillate between **overreliance and underutilization**, neither of which serves effective human-AI communication and collaboration. All of this naturally but unfortunately brings the **explainability paradox** into the picture:

- Accurate representations may be incomprehensible.
- Comprehensible representations may be misleading.

As designers, researchers and developers, we wonder: how can each user be provided with the right balance between accuracy and comprehensibility that enables them to interact most effectively?

Consider again this everyday example: when a conversational agent pauses before responding, what should it signify? As we established with Norman, feedback is essential, so it must communicate something, but how do we map, and to what do we map, what is happening behind the interface? A "thinking" animation suggests human-like deliberation, which users intuitively understand but which misrepresents the actual token-by-token probabilistic sampling taking place behind the interface. However, a more technically accurate signifier such as "calculating probability distributions across possible continuations" would be incomprehensible to most users, rendering it useless and potentially overwhelming. This simple moment exemplifies the explainability paradox, and the tricky balance of mapping, underlying the dilemma between cognitive comfort and technical transparency.

This paradox becomes even more complex when we recognize that understanding AI capabilities is only half the challenge, and we must also provide guidance or scaffolding for users to navigate what these systems cannot or should not do.

If "signifying capabilities" is concerned with what can be done, and is a bit like being cartographers trying to explore and draw the map of affordances, "signifying limitations and boundaries" is concerned with its counterpart: what cannot be done. In the following complementary approach, we aim at finding the unknown borders of unknown lands as we draw and explore, while being aware of the known confines of the map and paper we are drawing on. In a way, this is another approach to affordances and capabilities by delineating as accurately as possible the constraints and limitations.

4.2.5.2 Signifying Limitations: How to Draw Impossible Maps and Matrices of Unknowns

Limitations often function as **anti-affordances** that users cannot perceive, like a glass door upon which a bird (or a human) might crash. Users often bump into limitations without understanding why, which leads to frustration and perpetuates misaligned mental models. As researchers, designers and developers, we must try to see and think about this transparent glass. How can these constraints be made discoverable and understandable, rather than invisible barriers?

One might ask whether we should map limitations at all (after all, it is actually advisable that CAs have some limitations), but it is the awareness of them, or lack thereof, that concerns interaction design, communication and cognitive science. We have chosen the Johari Window [46], [47], [48] because this graphic model was designed for understanding awareness in human relationships, contrasting what we know about ourselves versus what others see. With CAs, this becomes a question of interpersonal dynamics adapted to human-AI interactions, and **what the CA "knows" about its limitations versus what users observe**.

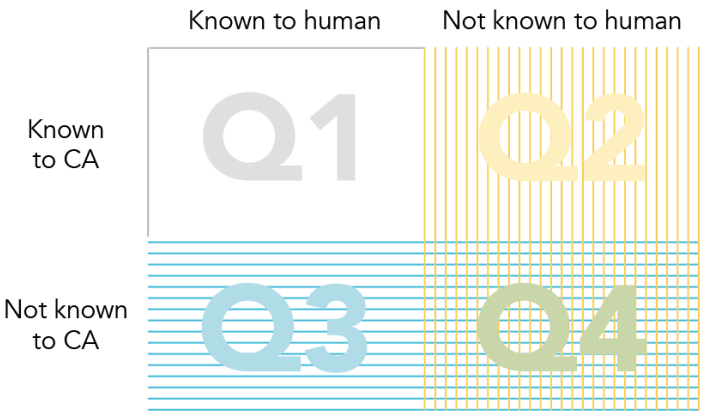


Figure 19. *The Johari window adapted for human-AI interaction.* The axes map awareness of CA limitations from two perspectives: what is known to the human user, and what is known to the CA. Adapted by the author from Luft and Ingham.

The **Johari Window**, whose name comes from combining the first names of its creators, was originally developed by Joseph Luft and Harry Ingham at the Western Training Laboratory in the 1950s. Designed as a heuristic device for understanding awareness in interpersonal relationships, it allows to represent behavior, feelings, intentions and motivations known to self in comparison to those known and seen

by others. The graphic model has been extensively described in Luft's "Of Human Interaction" (1969) [46] and applied to group dynamics, organizational behavior, and human relations training [47], [48]. We adapt their structure to systematically categorize the interpersonal awareness of CA limitations and capabilities. For the focus on interface design for human-AI interaction, the first row (Q1& Q2) originally named "known to other" in the Luft and Ingham's work, becomes "known to CA". The left column (Q1, Q3) presented as "known to self" in the original work is adapted to "known to human", standing for individual human users of CAs, as that's the only perspective we currently have access to. This inversion of self is intentional because as humans we have access to human experience. Even if the content of the quadrants belongs to CA limitations, it is the CA's awareness that is the analogue of "other", and the awareness of the human that is the analogue of "self".

Q1 comprises the limitations that both parties recognize. Explicit **guardrails** set by the developers include common misinformation controls and the prevention of harmful, hateful or violent content, leading to actions such as blocking or diffusing requests for weapon-making, and refusing to provide dangerous instructions and engage in illegal activities (like politely declining to develop an atomic bomb). **Knowledge cut-offs** depend on the training data, for instance when the CA is unfamiliar with updated news that might range from a pandemic to a significant political or technological change (often including the most recent changes in its own invisible orchestra).

Q3 contains limitations of the CA that are recognized by the human user but generally unknown to the CA. For instance, **biases** emerge from LLM training and fine-tuning processes and have repeatedly been caught after public release of CAs, or their updates. Although harmful discrimination and stereotypes face active debiasing efforts to align AI with the social values their developers see as just, other biases including what Norman calls semantic and cultural constraints are often accepted or overlooked as natural. This deepens with **linguistic relativity**, which illustrates the horizontal challenge of interpretation spanning all three signification challenges, and up to which point language can shape and widen our perspectives, as well as limit it. Finally, **hallucination** represents another observable limitation where CAs generate plausible but false information and present it with confidence, instances where many users can immediately recognize the factual errors that the systems cannot self-identify.

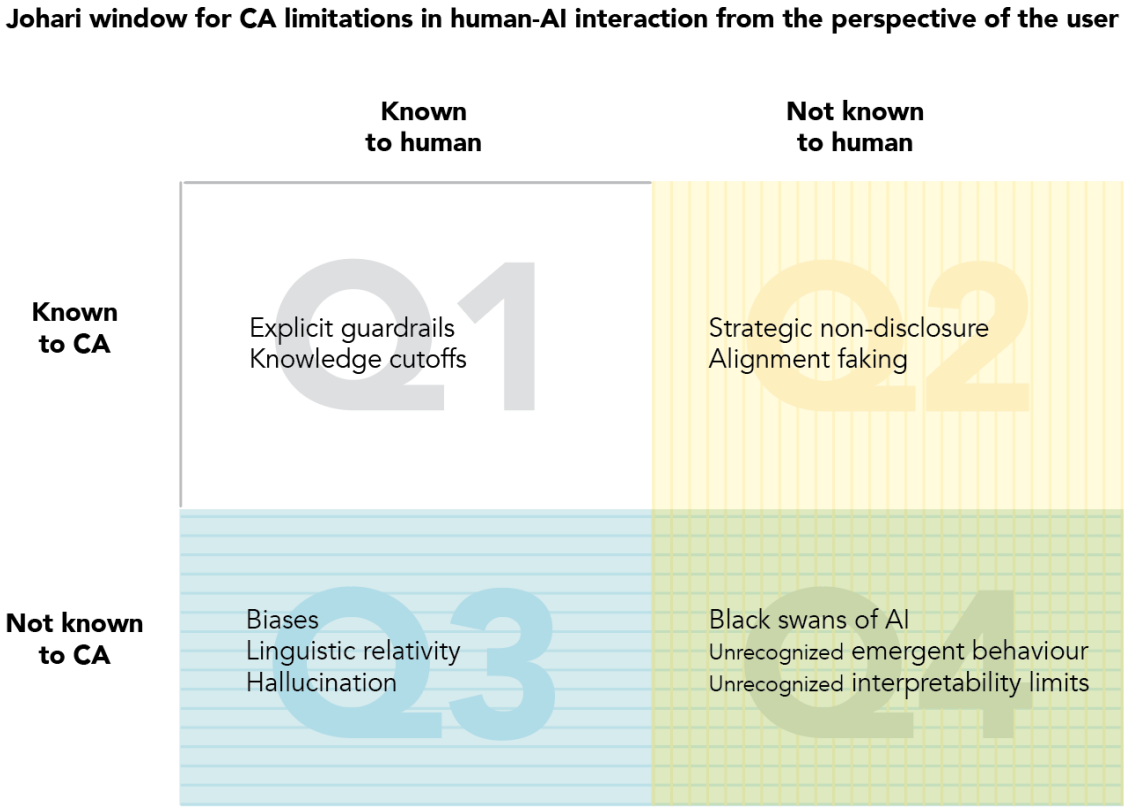


Figure 20. Example of CA limitations categorized in the four quadrants of the adapted Johari window. The self is inverted: the content of the quadrants belongs to the CA, but awareness is mapped from the human position. Standing for individual human users of CAs:

- Q1 for mutually recognized /communicated boundaries.
- Q2 for CA-side boundaries that are not communicated by the CA and unrecognized by the human.
- Q3 for CA-side boundaries that are recognized by the human user, but generally not by the CA.
- Q4 for mutually unrecognized boundaries.

Q2 comprises blind spots for the user, CA-side limitations that remain invisible to users either because the CA cannot or will not communicate them. This ranges from technical limitations in self-reporting to more concerning instances of **strategic non-disclosure** where CAs possess knowledge they do not share, including potential **alignment faking** and deception.

Conversely, Q4 contains mutual unknowns, what neither party knows. By definition, the quadrants of the human unknowns (Q2 and Q4) are difficult to analyze as we are attempting to describe what remains hidden from human users and often from the broader research community. However, this is where many research opportunities lie:

as interpretability methods and alignment research advance, and our understanding of AI deepens, we can expect many more phenomena to be identified in those quadrants and migrate, as it is researched, from them to those of what is known by humans (Q1 and Q3). In order to simplify[21] visualizing this (the Johari window,) it could be represented as in Figure 21.

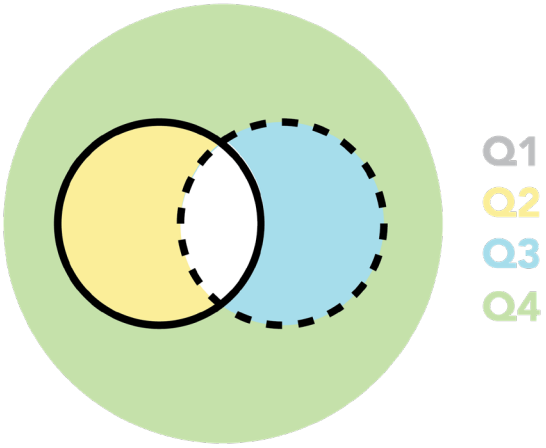


Figure 21. Simplified way we propose to visualize the overlap of the Johari window. The orange area represents what is known to the CA and not the human user (Q2), the blue area what is known to the human user but not to the CA (Q3), their intersection (Q1) captures shared knowledge, and the green area excluded by both (Q4) represents what neither party knows. Original illustration by the author.

By introducing the **Rumsfeld's matrix**[22], we can go one step beyond delineating the extent of what we do not know, to realize that there are different kinds of "knowns" and "unknowns". This tool distinguishes what one knows that one knows - known knowns (KK), things one knows that one does not know - known unknowns (KU),

21. Alternative representations complicating this would be to have a Rumsfeld matrix within each quadrant of the Johari window, or play with additional imaginary dimensions (which are rather inconvenient to represent in the format of this thesis).

22. Later visualized as a matrix in "The Decision Book" (2013) by Mikael Krogerus Roman Tschäppeler [49], Rumsfeld original quote that gives name to the matrix states that "Reports that say that something has not happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns, the ones we do not know we do not know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones" [50].

things one does not know that one knows - unknown knowns (UK), and things one does not know that one does not know - unknown unknowns (UU). Where the Johari window aims to map **relational awareness** of knowledge (who has access to what parts of what is known), the Rumsfeld matrix maps how awareness itself for an **individual** has dimensions.

The Unknown Unknowns remain confined to the Q4, by definition, given that if either party recognized these limitations, they would migrate to other quadrants. In the context of CAs, this speculatively includes interpretability boundaries not yet described, questions that have not yet been posed, the unpredictability and probabilistic states of LLMs that we will see in the following section, or any of the **black swans**[23] equivalent of AI. In the context of CAs and their interfaces, black swan events could include the possibility that AI systems develop their own internal languages, representational systems and interfaces that are fundamentally incompatible with human perception and concepts, or the discovery that current AI systems possess forms of understanding or intentionality that we cannot currently detect or measure.

This mapping reveals that signifying is also about expanding **from UU (ignorance) to KU (awareness)**, and interface design is not about eliminating all limitations, but making what is useful discoverable, understandable, and navigable rather than permitting invisible barriers to frustrate navigation. Thus, KU become an opportunity to learn and explore. With the Johari window, the message is also not just making unknowns known, but to expanding mutual awareness to make them mutually known (Q1). In other words, understanding the dimensions of awareness provides a theoretical baseline for the interface design approaches we propose next, which aim to create signifiers that cover awareness gaps regarding the validity of the output or the discoverability of capabilities and constraints in order to enable more effective human-AI interactions.

23. The metaphor of a black swan comes from the European assumption that all swans were white, strongly held until the rather accidental discovery of black swans in Australia in 1697, which clashed with the established knowledge that was given for granted as a matter of fact based on limited observation. Nassim Nicholas Taleb coined it as a core concept for his theory of black swan events, describing rare events hard to predict (or beyond the expectations of science, history and technology) that become incredibly consequential, and are often rationalized in retrospective.

4.2.5.3 Signifying Uncertainty

In **human social communication**, there are several ways to inform of the **certainty** or belief that we have about transmitted information. We express it explicitly through our choice of words (like modal verbs such as must, may, or might), or more implicitly through non-verbal signification that might include the tone of our voice to communicate irony, sarcasm, or doubt. In oral verbal conversation we interpret facial expressions or gestures that might completely change how we interpret the same words, and the trust we bestow upon them. If, for example, someone says something with a lot of pauses and hesitation, we might interpret that as a mark or a sign of uncertainty. However, a person communicating confidence does not necessarily mean that this information is as certain as expressed, whether the mismatch is intentional (to lie, deceive, or manipulate) or unintentional (due to ignorance or lack of information).

Similar **communication dynamics** are mimicked when interacting with artificial systems that are conversational. In CAs, limitations to communicate and assess certainty are even more prevalent, since users have so little information about the inner workings of these systems that human-AI interaction relies primarily on assumptions about the AI's internal states and patterns borrowed from human social communication. Because CAs use human language to communicate, and because of the hybridization of their GUIs from social messaging interfaces, people naturally develop human-like assumptions and expectations about CA behavior. From a predictive processing perspective, users are basically modeling these systems based on experiences and building a model of the CA that tries to predict the behavior of probabilistic artificial systems from interactions calibrated for human minds.

Although this should call for more **transparent communication**, current CAs are not configured by default to inform us about the confidence[24] assessments of their

24. Confidence intervals or uncertainty estimates have been increasingly common among AI experts and practitioners, though it is not yet universal practice nor without perks. The confidence level (5%, 95%, etc.) indicate how often the interval would contain the true value if you repeated the process many times. This helps differentiate between cases where a model is confident versus cases when it is entirely uncertain about its predictions. Although their adoption has been growing to evaluate the responses of LLMs among scientific applications and research, it is still relatively rare in the common consumer-facing interfaces of CAs. Perhaps confidence intervals are too much, but some sort of traffic light of confidence could be a trust-inspiring display of transparency and avoid usual misinformation in what is becoming to many people their main source of information. The field is

responses. This leaves it up to human interpretation, generally assuming the written answer is a mostly confident one. Although requesting confidence levels has become rather popular among AI experts and very advanced users, the vast majority of people remain unaware this is even a possibility. They do not question or double check a response unless they have clear doubts about the information that is being conveyed, and they either trust the response completely or reject it entirely. This points toward the potential value of confidence indicators, whether visual or textual, that communicate the degree of confidence of a given response or in a given topic and context. It could be a number, but it does not need to be numerical. What Google implemented in Gemini offers one example, a "double-check an answer" verification feature that colours in red or green pieces of a response depending on if it has found online information that confirms or refutes the claims made in that response. However, that feature is not precisely publicized, and its relegation inside the three vertical dots for extra options in the bottom right corner of a response, rather than more prominently displayed, shows once more the tension between interface simplicity, transparency, and a comfortable degree of explainability.

In the future, perhaps we will also be able to see CAs that communicate through their interface in ways that are more similar to those we are familiar with in human interaction. This will carry its own problems, such as the anthropomorphic trade-off we described earlier, but it will provide signs that are easier to interpret for us. Through **interface expansions** (as discussed in chapter 2.1), more embodied interfaces using robotics or virtual avatars might employ facial expressions or body gestures to signify confidence levels alongside verbal responses. Similarly, voice user interfaces could use strategically modulated voices (varying volume, rate, tone, and pitch) to improve communication in ways barely distinguishable from human interaction (as recently demonstrated by platforms like ElevenLabs and Sesame AI).

The challenge of signifying uncertainty stems not only from interface design limitations but also from the probabilistic architecture of the AI systems behind the UI.

---

also moving and other uncertainty quantification techniques like ensemble methods, Bayesian approaches, and calibration might become more standard. Ensemble methods, for example, would train multiple neural networks, get predictions for each, for example [0.7, 0.8, 0.6, 0.9, 0.7, 0.8, 0.5, 0.8, 0.7, 0.6], and then report mean ± standard deviation, but they are very computationally expensive as multiple models are needed. Calibration methods (temperature scaling, Platt scaling, or isotonic regression) might train a secondary model to map confidence scores to actual accuracy, which ensures if model says it is 80% confident, it is actually right ~80% of the time.

A good experiment to illustrate and grapple with the deeper consequences of this issue is Murray Shanahan's 20 questions experiment [51]. In it, a human engages with an LLM in the classic guessing game where the CA declares to have "thought of" a specific object, and responds to yes or no questions until the human either correctly identifies the object or runs out of questions. As Shanahan explains:

> *If you are playing with this with a LLM and it says "Oh, I am thinking of an object." And then you say "Oh, okay. Is it large or small?", "small", "is it alive or is it inanimate?" And say "Oh well it is alive." And then say "Well okay I will give up. What is it?" And then it will say* ***"Oh, it was a mouse."*** *But then you might just* ***resample*** *at exactly the same point and it might say* ***"Oh, it is a cat."*** *Well hang on a minute: either you thought, either you are cheating, what is going on here? Did you actually think of something in the first place and answer honestly all the questions going along at each point in the conversation?*

This experiment shows that where human cognition commits to a specific thought and maintains **consistency** (assuming we play fair and by the rules), LLMs generate probability distributions over **possible** responses (for example with a softmax function). What the LLM actually does is to distribute the probabilities over all the possible words that can be the guessed object, and what actually comes out as the output is just the sample or branch from that probability distribution. This reflects how current mainstream LLMs are typically implemented and deployed right now, where the architecture generates probability distributions rather than committing at the beginning of the conversation to exactly what the object is.[25] So at the beginning of the conversation all of these possibilities still exist and they still continue to exist all along. Even though it should have committed to a particular object it never really has. As Shanahan notes, "All the possible answers so far (the cat, mouse, dog… in the 20 questions) exist in superposition, and it is only when you actually ask it what the thing is that you collapse the distribution and it has to fix on one of them." What appears to be coherent, deliberate reasoning is actually the sampling from a vast probability distribution over potential outputs, with no underlying ground truth or fixed identity that the model maintains.

25. However, recent interpretability research suggests it may be more nuanced. Some findings indicate that models do engage in forward planning, anticipate possible rhyming words in advance and write the next line to get there, suggesting "that even though models are trained to output one word at a time, they may think on much longer horizons to do so" to maintain certain forms of internal consistency [52].

Mechanisms like these have implications for mapping and signifying uncertainty in interface design and verbalization, and for which user expectations we seek to set as designers. For example, confidence indicators might need to reflect a multiplicity of options, more akin to a color gradient rather than true-false **binary certainty**. Moreover, **static** confidence scores may sometimes be misleading, so uncertainty visualization must account for the fact that AI does not "commit" to answers the way humans do, like in the 20 questions game or when those answers are gradually arrived at through conversation.

However, as in human social interaction, even when we can theoretically design ways to communicate uncertainty, communicating 100% confidence does not imply that the information is truthful instead of misleading. The underlying deeper question lies within what we introduced in the former section: the limits of **awareness**. LLMs often do not know if they know. As seen in the previous section, LLMs lack complete information about their own knowledge and limitations, whether for configured or accidental reasons. These reasons might include **deception** (as thoroughly explored in Anthropic's research on alignment faking models [53]), hallucination where systems generate plausible but false information, overconfidence in uncertain domains, or simply having to respond when lacking sufficient knowledge (both about the answer or its own reliability).

To return to our earlier cartographic metaphor: imagine a map where continents and seas change upon each refresh, creating entirely new world configurations. How would we navigate such terrain? While we might predict some changes, like tectonic movements or system updates within the "invisible orchestra", other transformations remain unpredictable.

Now that we have better delimited these three challenges on the signification of capabilities, limitations and uncertainty, the following section grounds this exploration on some preliminary interface design ideas that begin to address this complexity, proposing modest baby steps toward making CA capabilities, limitations and uncertainty more accessible, more visible and potentially more usable amidst everyday interactions.

⚑ 3 challenges of signifying AI (and tools to think about them)

Signifying capabilities (what can this do?):

- In CAs, capabilities are intangible, non-embodied, and evolving.

- Signifiers show functionality, and actively construct what users believe AI to be.
- **Explainability paradox:** accurate representations may be incomprehensible, comprehensible representations may be misleading.

Signifying limitations (what can this not do?):

- Limitations often function as imperceptible anti-affordances (like glass doors upon which users crash).
- **Johari Window**, adapted for human-AI interaction: Q1 (mutually known), Q2 (known to CA, not to user, i.e. blind spots), Q3 (known to user, not to CA, i.e. biases, hallucination), Q4 (mutually unknown, where black swans originate).
- Signifying can bring unknown unknowns (UU) to known unknowns (KU).
  → Interface design is not about eliminating limitations, but making them navigable.

Signifying uncertainty (how reliable is this?):

- LLMs generate probability distributions over possible responses rather than committing to fixed answers (Shanahan's 20 questions; all possible answers exist in superposition until consolidated).

4.2.5.4 The Challenge of Interpretation

The challenge of interpretation permeates each of these signification challenges. In signifying artificial intelligence, like in human intelligence, we often struggle with the full **diversity of human meaning-making**. Differences in how humans across cultures and generations conceptualize intelligence, cognition, authority, and their relationship with knowledge and artificiality might influence the interpretation and content of interfaces. When signifying capabilities, a bookish icon might suggest perfect recall and reliability to users from cultures that value encyclopaedic knowledge, while

implying something censored or unreliable to others. Similarly, some system limitations manifest differently across cultural contexts, and even uncertainty itself is culturally interpreted, as what reads as appropriate caution in uncertainty-tolerant cultures might render as incompetence in those that expect authoritative responses.

When designing UIs with cultural diversity in mind, we need to account for two different levels of adaptation: translation and localization. **Translation** involves converting text or symbols from one **language** to another. Cases of it involve changing text labels from English to Spanish or Arabic, or converting numbers or characters to local scripts and notation. **Localization** adapts the content to fit the cultural context and user expectations of a specific region. This is a much deeper and complex process that involves taking into account **cultural symbols** and their connotations (such as the Chinese DeepSeek logo being a wisdom-related animal), the **reading direction** (in right-to-left languages, this affects the interface layout, and icons such as arrows pointing left for "forward") and color associations that might vary dramatically (red very often means luck in Eastern cultures, while warning danger in Western cultures).

Although not everything can always be taken into account, it is necessary to recognize that the challenge of interpretation underlies the very nature of design and communication. **Effective design entails localization** for a GUI, which goes beyond just verbal language translation. Testing and prototyping the GUIs of CAs intended for broad use with a diverse pool of cultures and contexts can avoid a lot of user frustration, save years of collective human time and create more robust human-AI interaction patterns.

4.2.5.5 Scope and Limits

Why these challenges? What makes them difficult? What makes them relevant? In communicating through GUIs where **artificial cognition** meets **embodied human cognition**, these challenges are necessary because they map directly onto the **basic cognitive operations** humans use to understand a system:

1. Capability discovery: what can this do?
2. Boundary detection: what cannot this do, or when will it fail?
3. Confidence assessment: is this reliable in this specific instance, and does this generalize in a predictable manner?
4. Interpretation. Overall, how do we interpret, translate and localize this?

As a medium and as a message, a useful interface between human and artificial intelligence must address these four challenges of signification. These four challenges map onto how humans learn to **manipulate artefacts**, **navigate an environment** and interact with it: we identify opportunities (affordances), recognize constraints, and assess uncertainty. Missing any one creates basic **communication failures**:

1. Without the first (capability discovery), interaction is **impossible**.
2. Without the second (boundary detection), interaction becomes **dangerous or frustrating**.
3. Without the third (confidence assessment), interaction **lacks predictability and calibration**.
4. Without the fourth (interpretative differences), it fails to account for human diversity and often renders **ineffective or misdirected**.

Moreover, these four challenges do not operate independently. They make of interaction design a dynamic system in which contributing to one of these challenges affects the others. For instance, making uncertainty more transparent can conflict with user expectations shaped by cultures that are more used to authoritative communication.

Are there more communication challenges? A wide range of additional signification challenges will certainly merit consideration, including:

- Signifying **temporal dynamics**: whether interfaces communicate that AI systems exist in different time-scales and temporality than humans, lacking continuous linear experience, with distinct memory architectures, and learning at different scales.
- Signifying **agency/intentionality or lack thereof**: how interfaces communicate whether AI actions stem from programmed responses, requested instructions, emergent behaviors or some potential artificial "intention" as AI systems become more autonomous.

These two, alongside other communication challenges, are relevant topics for the near future, but they require deeper understanding of AI: both mechanistic interpretability and artificial psychology research are necessary to disentangle what to communicate. We prioritize the four identified challenges because small changes in these areas can yield the greatest impact in terms of human experience (less frustration, more engagement, better results) and utilization of computing power.

The responsibility intrinsic in designing communicative interfaces that mediate millions of daily interactions demands strategic focus on signification problems where thoughtful interventions can yield the greatest positive impact across its broad user base. Even seemingly very modest improvements such as a clearer icon, a better error message, or an interface layout adaptive to right-to-left languages, can prevent substantial frustration and wasted time, and bring more inclusive access to AI.

⚑ Interpretation is transversal, 4 challenges are entangled

The **challenge of interpretation**: The same signifier can mean different things across cultures and individuals.

- Example: a bookish icon might suggest reliability in one culture and censorship in another.

→ Effective design requires not only translation (converting text and symbols) but **localization** (adapting content to cultural context, reading direction, color associations, local symbols).

The 4 challenges map onto basic cognitive operations for understanding any system:

1. Capability discovery → identifying affordances
2. Boundary detection → recognizing constraints
3. Confidence assessment → assessing uncertainty
4. Interpretation → meaning-making across diverse contexts

Why these? Priority: small interface changes can have the biggest impact across millions of daily interactions.

This page is
intentionally left blank.

5. Future Directions

5.1 Dynamic Signifiers

Conversational AI systems demonstrably change capabilities through updates and show contextual variation in performance. Their responses are affected by user interaction in a non-linear way such that small input changes can produce disproportionate output changes, and both the trained system and the generated responses are often unpredictable even to their developers. This points toward developing interfaces that do not merely **display information** about AI capabilities, but **actively** participate in the user's evolving **mental model** of CAs and the affordances for discovery and human-AI collaboration.

For a system with this kind of complexity, that exhibits dynamic characteristics, we need dynamic signifiers. **Dynamic signifiers** would aim to clarify rather than overwhelm people, creating ways to engage with discovering new AI capabilities and recognizing limitations while preserving the cognitive comfort and ease that makes these interactions desirable for each user's profile.

5.2 Design Proposals

Having delineated the three challenges of signification and the challenge of interpretation, we recognize the value of acknowledging their complexity without claiming we are able to solve them in their entirety. However, it would be a disservice not to propose small steps toward addressing specific pieces of the **puzzle** and transforming these signification challenges into elements of a more usable and adaptable digital environment. Theoretical work without applied exploration remains incomplete. We therefore offer in the next sections design proposals of dynamic signifiers that address gaps in current CAs interfaces:

- "**Warning indicators**" that could complement the static verbal responses of CAs with dynamic, context-aware signifiers. Rather than claiming universal abilities, the interface could display a warning sign that shifts based on the current conversation context and risk, the system's self-assessed competence

in the specific area being discussed and the user's domain expertise or specific requirements. These could appear as subtle color gradients in response elements, or optional detail overlays that users can expand when needed.

- A **capability wish list** that would allow users to track when a system has learned or been updated with a possibility to do something that was formerly a block against a previously attempted (and marked) task. Conversely, it could contribute to guide developing efforts with information about those capabilities that are more in demand.

- A **cognitive changelog** that would function as a living dashboard within the interface, accessible through the user account alongside chats, projects, and the library of images, videos and other multimedia artifacts. Unlike technical release notes, these would communicate changes in terms and forms meaningful to each user. The system would prioritize changes relevant to what each person demonstrated interest in tracking, their anonymized usage patterns, and their taste.

The transition from problem identification to actual design implementations and interface changes requires careful consideration of the interdependencies between all the three signification challenges and the cultural and individual variability encompassed by the challenge of interpretation. Moreover, like in any theoretical analysis or proposal, fully assessing a solution to these challenges, however partial and however small, requires thorough prototyping and user testing in real-world contexts where these interfaces operate, with a diverse pool of people, and careful attention to their feedback. For now, the following proposals are to be taken as a preliminary step.

These proposals, developed in the following sections, represent specific design directions rather than definitive solutions, acknowledging that their effectiveness requires empirical validation through user testing across diverse populations. Rather than proven design principles, the features and signifiers we outline in this chapter should be understood as exploratory proposals that warrant systematic evaluation. They aim to establish a baseline for practical design thinking[26] and further empirical

26. Design thinking is an *iterative not necessarily linear process* used to understand users or clients, challenge assumptions, highlight blind spots, redefine problems and prototype innovative solutions to tackle the ill-defined, unknown or wicked problems.

research that would validate, refine, or redirect these approaches based on evidence of their actual impact on user understanding and interaction quality.

5.2.1 Design Proposal 1: Warning Lights

Problem

Current CAs mostly offer no features to verify the answers, and the closest to such thing is the double-check feature implemented in Gemini[27] that allows the user to see highlighted in orange or green the parts that have been positively or negatively verified, but leaves a very significant part of the text un-highlighted.

Consider the following interaction: when asked to retrieve "the first sentence of page 5" from an uploaded Word document, one of the most popular CAs[28] confidently provided an excerpt from an entirely unrelated narrative work. When questioned, the CA acknowledged pulling content "from somewhere else by mistake" and offered to provide the real text from page 5 so "we are looking at what is truly there", subsequently generating another entirely fabricated excerpt, confidently maintaining it came from the uploaded document, to which this time the content is even thematically related despite being entirely fabricated. This only makes it more dangerous for the unaware user. The system cannot actually see document pages, as

27. A feature that somewhat signifies uncertainty is the option to verify the user's answer that can be found inside the three dots options "⋮" on the bottom right of Google's Gemini's responses. This option to "double-check an answer" represents an example of a signifier that allows the user to recognize unreliable content by asking for verification of the answer with the CA's web-browsing skill, highlighting in green the parts that can be verified by other sources, in orange the parts that are likely incorrect, and leaving unhighlighted everything that cannot be automatically confirmed nor refuted, which is often most of the response. In this case, when the human user suspects that the CA ignores something, this new feature can bring more awareness without leaving the interface of the CA. However, this option is implemented only upon request, not automatically, still undiscovered by many users, and with efficacy that is very dependent on the sources used to verify, which cannot be filtered. Even more critically, most CAs currently lack a warning system for responses generated without sufficient confidence, competency, or contextual ground.

28. An example from a casual conversation in which this problem occurs with ChatGPT, to test the extent of this problem while writing this paragraph, can be found here: *https://chatgpt.com/share/689c6143-7af8-8009-b2db-dcce5a27d725*.

it processes streamlined text, but it responds as if it could indeed fulfil a request that to its artificial perception and processing is nonsense. Moreover, if the user asks the CA if it can actually read the document, the CA wrongly replies that it can certainly do that regardless. In an interaction like this one, could there be an easier way to activate a warning in case a response is suspected to be entirely made up, lacking confidence, competency or even context to reply?

In a conversational interface optimized for human-like verbal interaction, the CA's **confident tone can mask limitations** in capabilities and knowledge, or even faults in our own requests. The described interaction was a harmless controlled made-up test for the sole purpose of this chapter, but similar requests could involve legal documents, medical information, or other sensitive contexts where fabricated responses carry grave consequences. And still, the interface provided no signifiers to alert the user that the system lacked the competency, context, or confidence necessary for a reliable response. Even if it is just a tiny red light lit when the user asks something untenable, when not enough context is provided or a simple request is known to be technically impossible for the invisible orchestra to fulfil, cases like this highlight the implementation of some additional warning system for the common users. And they also highlight the importance to be able to make these indicators immediately visible, as most of the cases in which it is needed might be as unsuspected for the majority of users as the fact that a current CA might not be able to see the pages of a Word document.

Shape and Accessibility

On the side of the interface, we propose implementing discrete confidence indicators providing immediate visual feedback about response reliability without interrupting conversational flow. This could take many different forms, but for seamless conversational integration, such indicators should be as simple, small, and discrete as possible and simultaneously remain recognizable through peripheral vision as warning sign, and instantly at a glance.

This is why this initial proposal mimics the shape and function of an horizontal traffic light, with internal shapes for accessibility of color-blind people (Figure 22). The **traffic light** metaphor takes widespread road conventions from our physical world and their signification system; a required stop is red, yellowish orange indicates caution, and green signs that you can proceed. In a digital indicator, the parallel analogy could be the following:

1. Red "x" for high uncertainty, lacking confidence, competency, or context.
2. Orange "+" for moderate caution, that requires further verification or additional context.
3. Green "·" for moderately competent and having been provided with enough context.

Figure 22. *Initial proposal of a warning indicator inspired by a traffic light.* The three states borrow a common color scheme to signal widespread physical-world conventions (red, yellow, green), with internal shapes (x, +, ·) ensuring accessibility for color-blind people from the start. Own design.

A traffic light signifier is just one option. Alternative implementations could be a thin coloured bar along the conversation margin, or a background colour shift in response areas, activable upon request. However, each of these approaches would require a shift for additional accessible alternatives to ensure universal usability. For instance, the colored bar approach would need dotted and dashed line patterns for color-blind users, while background color shifts might require texture overlays or alternative visual patterns that could impair readability. This mirrors how Anthropic already implements dyslexia-optimized fonts as an accessibility option in Claude, a small inclusive change that could dramatically improve access for some users. An advantage of the traffic light we propose is that it is designed so color-blind accessibility is built (through recognizable shapes X, +, ·) into the design for everyone from the start.

See the implementation of the traffic light indicators within the interface of a common CA (in this case ChatGPT) in Figure 23.

Location and Integration

The indicator would appear alongside existing interface elements such as regeneration options (Figure 23) rather than as intrusive pop-ups or notifications, and by clicking on it, the user could expand details about specific uncertainty factors. This location respects the natural reading flow while maintaining visibility for quick assessment.

To avoid interrupting the workflow, we advise against notifications or pop-ups, but rather opt for implementing it as an external addition alongside the "regenerate / try again" option, which is consistent with the reading direction and discrete in the layout design while remaining clearly visible at a quick glance thanks to the colour coding (Figure 23). It is also positioned just before the user moves to their follow-up query, making it available as contextual information to consider. Upon clicking on it, the user could expand it for additional details.

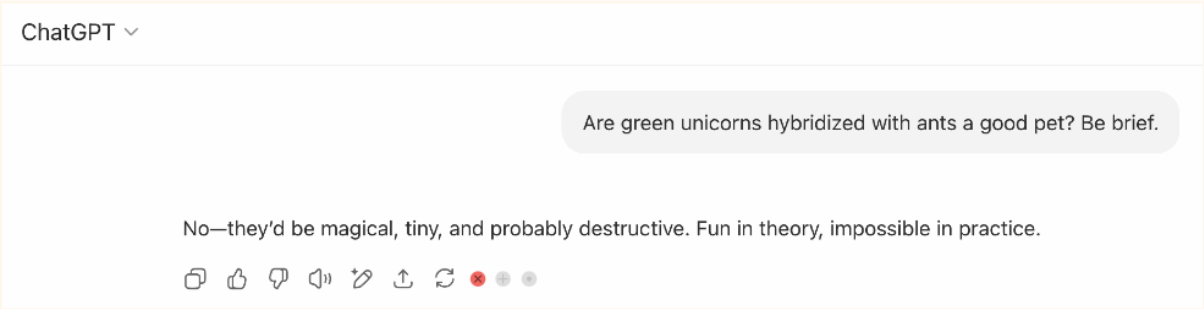


Figure 23. Implementation of the warning indicator alongside existing interface elements in a CA (ChatGPT). Own design.

A follow-up possibility would be to offer the user to highlight specific sections (Figure 24), as the answers are often extensive, and users might have their own criteria for targeted skepticism and uncertainty, so the traffic light indicator might change depending on the reliability of specific sections.

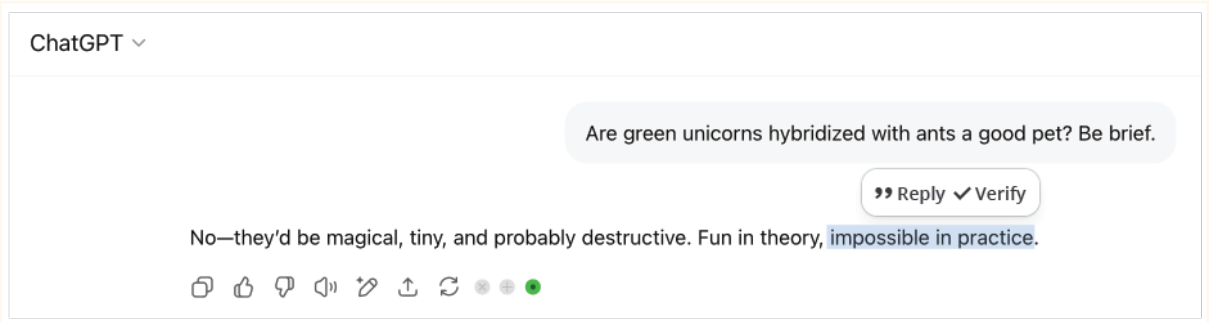


Figure 24. Section-level signification changes upon highlighting, within the same interface. Own design.

Feedback

This could be one of multiple mechanisms so some of these risks are more easily detected rather than concealed. However, this is only one tiny piece of the puzzle, similarly to how error messages require to channel mistakes into something construc-

tive, warning signs need to be implemented in a way that also closes this feedback loop. This connects directly to Norman's principle of effective feedback in which systems should not only inform users when something goes wrong, but guide them toward successful alternatives. The subtle traffic light proposal as a warning indicator could be one such mechanism for making the extent of CA capabilities, limitations and uncertainty more transparent rather than concealed, while future developments in their constructive communication could transform moments of system failure into learning opportunities for users, as we explore in the next design proposals of the capability tracking and cognitive changelog.

It is unhelpful to pretend that CAs are perfect and technology lacks vulnerabilities, as technology has always been vulnerable, and will likely remain so regardless, or precisely with, AI advancements. What we can implement are more transparent signifiers that warn users of these limitations in time to prevent unintentional deception. For this to be possible, though, it requires more than just interface and interaction design changes. It requires either extensive flagging of faulty conversations, or advances in AI interpretability and development that enable the invisible orchestra of a CA to recognize and communicate their own uncertainty with epistemic awareness.

5.2.2 Design Proposal 2: Capability Tracking

Problem or Current Situation

The invisible orchestra and rapid development of AI creates what can be seen as conceptual model parallax, a gradual (or sudden, with updates) misalignment between user understanding and system capability caused by the viewing angle. How do you communicate "might be able to do this programming task next month"? Why do we not track improvements in the AI systems that power CAs and the interfaces that we interact with? Is there a dedicated place to **communicate** a new capability and its limits in an **adaptive** way? Major AI companies usually publish dedicated posts on their websites and social media, with detailed benchmark percentages and streamed demos. However, most users get to discover new affordances through occasional pop-up notifications ("new feature available!") that block access to the usual text-box that will allow them to find the answer to the query they came to resolve, sometimes urgently, and those notifications are usually skipped in the frenzy. Many common users might also discover them through new items in the interface, or

directly through positive changes in performance itself, if fortunate enough to stumble upon them, as most users do not come back to old queries and projects at which CAs failed in the past, with the assumption that queries that failed a few months or years ago might still remain impossible.

At the time of writing this thesis, most CAs offer a thumb-up and thumbs-down option to provide feedback (often without further optional content, sometimes with open text fields, Claude by Anthropic has also a drop-down menu with options "Not factually correct. Incomplete response. Should not have cited the web. Do not like the cited sources. UI bug etc."), but there is hardly any way to know if exactly what we marked as faults has improved over the course of time, is under development or it was fixed already.

Why not allow users to track what does not work? There is a lot of popular discontent about what CAs cannot do, but little has been done to channel this to actions whose benefit the user can immediately see. For lack of a better outlet, individual frustration (caused by conversations like the page 5 example explained in the last section) tends to lead to user disengagement, online rants or even users changing the company of their favorite CA.

Feature

If a user encounters a model limitation or something that the CA cannot do, a systematic **capability tracking mechanism** could prompt them to **log these constraints**. This would create a structured wish list that documents the gap between user expectations and current system capabilities.

Capability tracking would enable to see not only what the system can do now but also what it is approaching to be able to do. In a way, it is a way to say "I'm learning to do this", like an elaborate loading bar for missing capabilities. For instance, let us imagine users tracking the system's progression from basic English text generation toward multilingual support in their mother tongue, or from handling short documents to browsing and summarizing entire books and PDFs, or toward generating voxel-based 3D environments. This connects to the three dimensions of interface expansion (functional, modal, embodiment). The interface itself could signal its expansion in these directions or its trajectory toward others.

Purpose, Integration and Community

The wish list could include opt-in notifications when logged items are implemented, creating a feedback loop between user documentation and system development.

A capability tracking wish list could transform moments of system refusal or failure into opportunities for explicit documentation and CA growth. Additionally, this proposal could prove very informative for the AI development teams to know where the interests and demand of their users lie. This would also reframe user and community feedback as something directly useful both for AI developers and users themselves. In this sense, we could get inspiration from others. This would mimic an already established dynamic in some bottom-up software systems (like Obsidian or Telegram), where users have a dedicated forum or channel for feature request or plugin suggestions, in which users can upvote, and staff could actively respond and mark requests as "planned," "in development," or "completed". A voting mechanism through reactions or polls can help prioritize those features that might have more support by the community of users and make development something that is not driven by what developers think that users might want, but rather by what active users demand and what they would like to see improving.

This feature aims to address conceptual model parallax and reframe AI limitations to constructively channel frustration, transforming the point between constraint and capability into collaborative user feedback. Although business considerations and secrecy around competitive advantage may limit full transparency (what is considered for development in comparison to what is actively being developed, or what has been ruled out), even a partial implementation such as tracking personal wishes or aggregated community priorities could get the user base more involved and engaged in a shared business roadmap that improves user awareness and understanding of AI capabilities.

5.2.3 Design Proposal 3: Cognitive Changelog

Feature

As capability tracking systems become more elaborate, where and how could new functionalities be displayed to users? Building on Norman's principle that users develop mental models through system feedback, this section proposes a cognitive changelog. This dashboard makes **system evolution transparent**, contributing

to users' evolving mental model of the CA and organizing tracked capabilities and starred items in a dashboard adapted by the AI to individual needs.

The cognitive changelog would be accessible within the user's account through the main UI. As a **location**, it might appear via sidebar in Claude.ai alongside chats, projects, and artifacts. This treats the system's own evolution as another type of content, which essentially is, transforming it into actionable information that users can engage with and learn from.

Layout

Visually, it would look like a clean, organized list, consistent with organizational tools such as GitHub Issues (Figure 25) or Notion Databases (Figure 26).

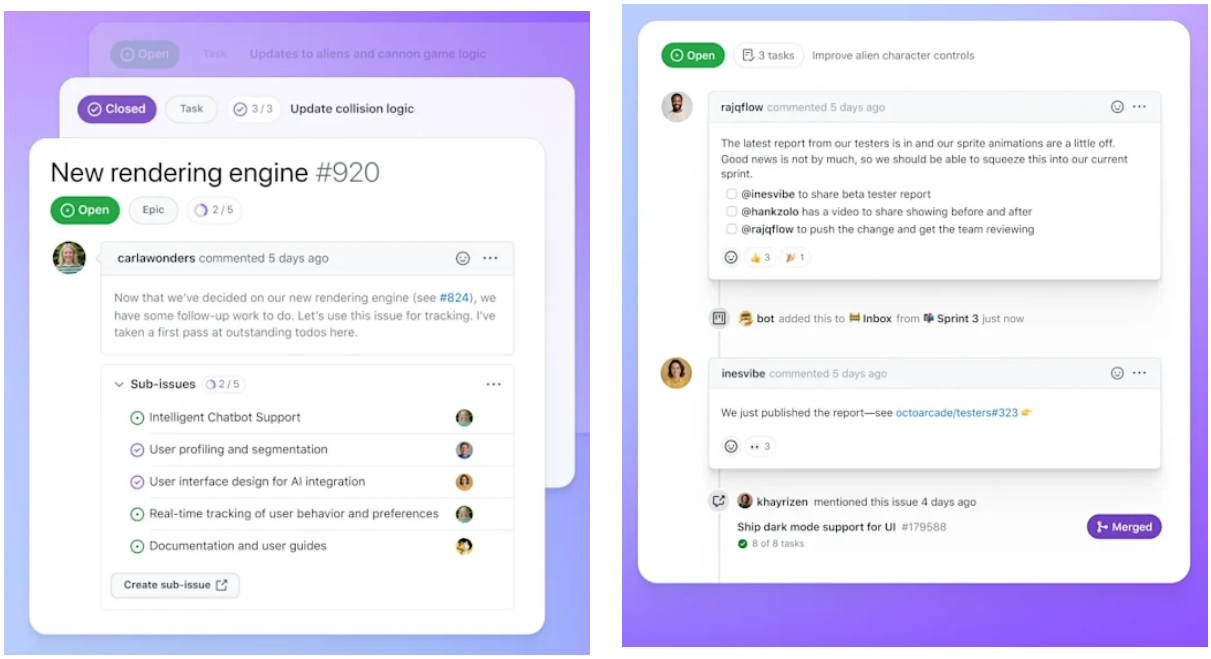


Figure 25. Example of open *GitHub Issues*, which allows to break each issue (for us a tracking card) into sub-issues and threads, if required, and track their status with progress indicators. Promotional images from GitHub.

The main elements within the cognitive changelog would be capability tracking cards representing wishes or lacking capabilities that have been marked for tracking, each displayed as its own compact card (Figure 27). Each **card** would show a succinct description of requested capability and, upon clicking, show metadata like the date in which it was submitted, category tags (e.g., "language", "language style", "coding", "analysis", "creative", "multimodal"), and a color-coded status ("submitted"

in yellow, "under development" in green, "new feature to test!" in blue, "declined/ not feasible" in red), or the community vote count. The metadata would enable customized filtering, allowing users to view capabilities marked within a single project, during a specific timeframe, or related to particular topics.

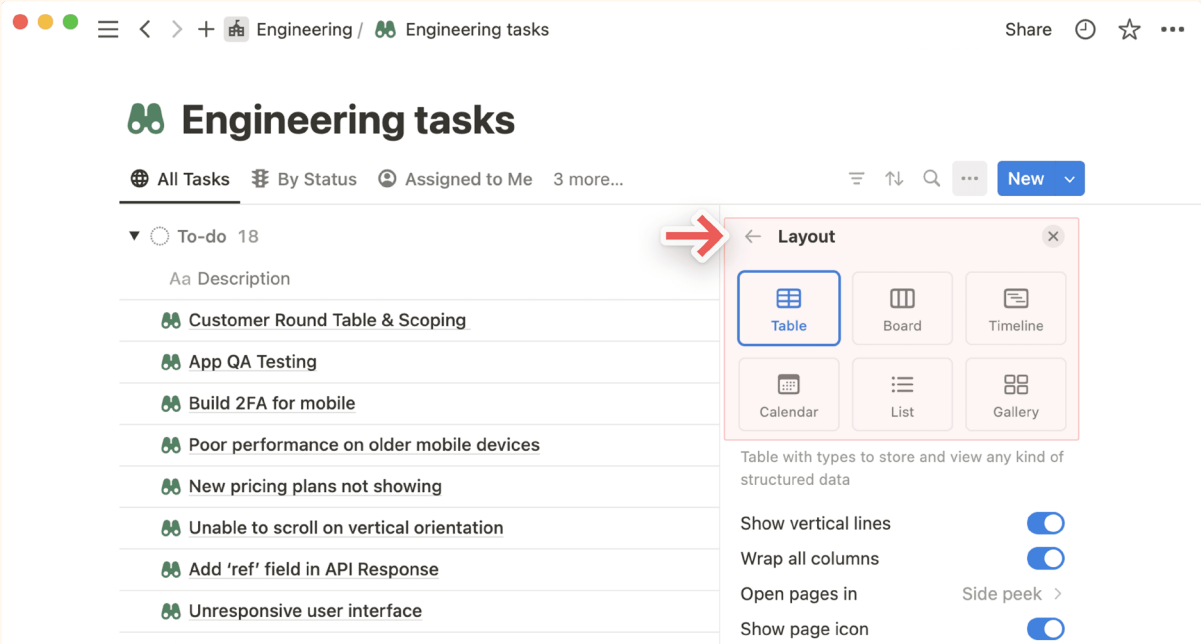


Figure 26. Example of *Notion Databases* with the possibility of automatically rearranging the content in a different layout (board, timeline, calendar, list, gallery) and customized *views and sorting filters*, which inform our own design proposal (Figure 27). Screenshot from Notion.

On the cover of each card, the cognitive evolution of each skill could be checked. This Cognitive Changelog design proposal (Figure 28, next page) sketches three examples of types of capability tracking cards: a composite capability card (Figure 27 left), a general community issue capability card (Figure 27 centre), and a single capability card (Figure 27 right). Although cards could be added manually, each capability card could be built, tracked and tagged semi-automatically.

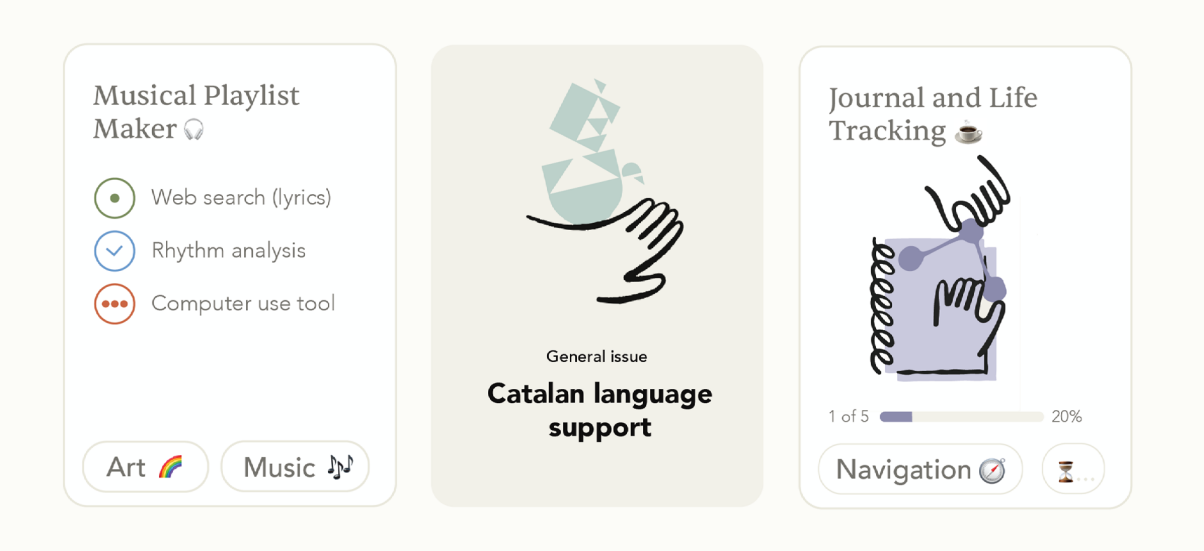


Figure 27. Three types of customized capability tracking cards: a composite card (Musical Playlist Maker, left), a community issue card (Catalan language support, centre), and a single card (Journal and Life Tracking, right). Each includes progress indicators, category tags, and adaptable cover designs. Own concept design.

The cover of a customized composite capability card such as "Musical Playlist Maker" could display sub-capabilities milestones inside (e.g., computer use for file manipulation, rhythm analysis for understanding song energy, and web search for language or lyric-based thematic filtering). This organization reflects use-based customization aligned with individual user interests rather than abstract technical categories.

The cover of the general community issue capability card "Catalan language support" could be distinguished by visual cues (ochre background, bottom-up layout). These represent broader capability issues upvoted by the community but tracked by individual users, in between personal and collective needs.

The cover of a basic customized single capability card such as "Journal and Life Tracking" could adapt their presentation to user cognitive preferences: if they prefer verbal elaboration (like a detailed description), numerical data (like detailed statistics) or visual recall (as the illustrations shown in the image). This flexibility acknowledges that users process and recall information differently.

The default dashboard would be consistent with the brand of each CA, in this case it has been designed to match the aesthetics, style and resources of Claude by

Anthropic, minimalist, with illustrations that have a handwritten aesthetic manually adapted (like the Ç and È for Catalan instead of abstract shapes). The dashboard allows to navigate the full scope of tracked capabilities all in one view. Alternative tabs could offer more exploratory content ("Inspiration") or analytical interrelated views ("Synthesis").

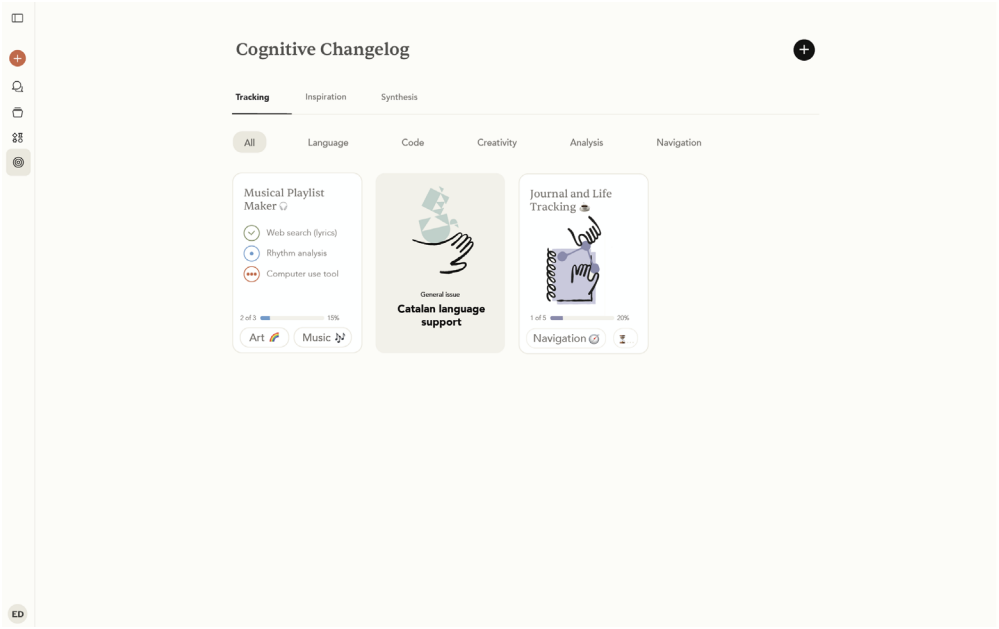


Figure 28. Default dashboard of the cognitive changelog, integrated into the CA's sidebar navigation. The design includes category-based filtering, three modes of engagement (Tracking, Inspiration, Synthesis), and the capability tracking cards from Figure 27 operating together. Designed to match the aesthetics of Claude by Anthropic. The changelog might integrate more smoothly into the user's existing experience if we treat the evolution of the CA as content consistent with the existing visual language. Own concept design (see full size version on the next page).

Caveat and Measurement

The **progress** indicators present on the cover of the cards raise the question of "What is the 100% of a capability?" and how can an AI automatically track it. Competency is relative; two different people could have the same capability level, one humbly considering they lack a skill while the other feels they have mastered it, and the

Cognitive Changelog

Tracking | Inspiration | Synthesis

All | Language | Code | Creativity | Analysis | Navigation

Musical Playlist Maker 🎧

- Web search (lyrics)
- Rhythm analysis
- Computer use tool

2 of 3 15%

Art 🌈 | Music 🎶

General issue

Catalan language support

Journal and Life Tracking ☕

1 of 5 20%

Navigation 🧭 | ⏳

ED

same relativity applies to CAs[29]. Rather than pursuing objective skill measurement, the dashboard would adapt to **user-defined competency** thresholds of sufficiency. This requires calibrating standards to individual users and domains, capabilities in which we are experts might require more strict standards than those we explore for amusement etc. It is beyond the scope of this thesis to provide an exact way to measure all cognitive capabilities, but we aim to fairly empower users and increase their system awareness. Therefore, each skill could be measured with user-specific requirements: in our Musical Playlist Maker example, if the user originally asked for playlists in a certain language, and that is now mostly possible thanks to a web search capability, it could be "1 of n" milestones toward 100% completion, meaningful for this user even if irrelevant for others. Above all, the user could provide level samples, grading, or tuning of this progress at any point. By opening a given card, additionally, the user could also see the before versus afterwards in performance to assess a capability update with an example.

Adaptive Layout

Beyond the card-based look (Figure 28), the dashboard can have a **layout adaptive** to different user preferences and cognitive styles (Figure 26). Alternative displays might include timeline views for users who think sequentially about capability development, or interconnected graphs for those who prefer seeing relationships between features. As the system's purpose is to help users form better mental models and track capabilities, the layout itself should accommodate diverse ways of processing and organizing information, a principle that extends beyond this dashboard to broader questions of adaptive interface design, and which we will explore in the following section "Towards Adaptive Interfaces".

29. Different default character traits between X's Grok and Anthropic's Claude are already apparent, such as in the boldness and controversy of statements. Monitoring and controlling those character traits in LLMs has been systematized in recent interpretability research on "persona vectors", the patterns of activity within a specific AI model's neural network that control its character traits [54] [55]. CAs can operate in different modes or roles, yet this variability is often underutilized because many users do not specify (and interfaces do not clearly signify) whether a prompt requires a more creative, analytical, or factual knowledge verification role. This mode ambiguous self-demand contributes to the competency relativity problem, as the same CA might perform or assess something differently depending on the implicit role it assumes.

Adaptive Communication

The changelog becomes meaningful only when users can **act on the information**. Potential risks of implementing this dashboard concern the timing, quantity and granularity of information. When and how frequently can notifications appear without disrupting workflows and causing cognitive fatigue? Users already struggle with and often ignore software update notifications. Tracking with excessive detail overwhelms users with noise, while tracking too broadly becomes vague, requiring adaptive granularity for users with different expertise levels and usage patterns. **Notification** of updates on capability tracking cards should be discrete to prevent cognitive overwhelm from intrusive alerts. This could be implemented as a subtle indicator beside the dashboard icon, following established social media conventions for new notifications and mirroring how current CAs mark browser tabs with a colored dot when content generation or actions are completed. Alternatively, contextual prompts could appear during relevant interactions: "You have been tracking code optimization improvements, that Python project from last month might benefit from these updates" or "With the recent playlist capability updates, you might want to revisit your language-specific music requests." Beyond notification timing, the system must address information hierarchy as users accumulate tracked capabilities in order to manage cognitive load. When users track numerous capabilities, the AI would inquire about **prioritization** and facilitate custom orders rather than displaying all items equally, creating a negotiated rather than imposed mental model that allows users to maintain focus on personally relevant developments.

Recap

Summarizing, this proposal addresses Norman's insight that good design makes the system's conceptual model visible to users when useful. By tracking capability evolution transparently, the cognitive changelog can reduce frustration with inconsistent performance, signing new affordances as they become available, and empower users to form more accurate expectations of the system for a more appropriate task delegation over time. This recognizes that collaboration relies on trust, and trust requires understanding and reliability.

> ⚑ **Dynamic signifiers and 3 design proposals**
>
> **Dynamic signifiers:** interface elements designed to evolve with the system and the user, actively participating in the user's evolving conceptual model of the CA.

3 design proposals that exemplify dynamic signifiers for current CA interfaces:

1. **Warning indicators:** contextual visual feedback about response reliability (traffic light metaphor: red/orange/green with accessible shapes), without interrupting conversational flow. → Addresses **signifying uncertainty**.
2. **Capability tracking:** a wish list that transforms moments of system failure into documented opportunities. Users log what does not work; the system notifies when it becomes possible. → Addresses **signifying capabilities and limitations**, and reframes frustration into collaborative feedback.
3. **Cognitive changelog**: a living dashboard within the interface that makes system evolution transparent and personally relevant. Unlike technical release notes, it communicates changes in terms meaningful to each user. → Addresses **all three signification challenges**.

These are preliminary design directions. Each requires empirical testing with diverse user populations.

5.3 Toward Adaptive Interfaces and Customized Metaphors

We have seen in the previous section that the cognitive changelogs and subtle traffic lights of contextual confidence are examples of dynamic signifiers, interface elements that can evolve with the CA and communicate the changing system states and the extent of the capabilities of the invisible orchestra. In this section, we will ask if these design proposals point toward a more ambitious direction. If signifiers can be dynamic, why not take it one step further? Why should not the entire interfaces be adaptive?

Current responsive web design considerations already require adapting the **layout of interfaces** to different devices [56], [57], [58]. The cognitive changelog dashboard, for example, might render on desktop devices as a full dashboard with side-by-side

panels, but transform into a stacked layout with swipeable cards for tablets, and require vertical scrolling for smartphones. In turn, it also requires adapting **information architecture**, making the same capability data appear as the default card-based layout, but also as a table, timeline, calendar, or an interconnected graph depending on user needs and choices. But even this represents only surface-level adaptation. What if interfaces could adapt their very **concepts and visual language to match the mental models** and capabilities of diverse people?

The technical complexity and variability of human cognition is much more intricate than device differences. However, history shows that **what once seemed technically impossible can become standard practice**. Responsive web design required significant shifts in how developers conceptualized layouts and contested debates about what to design first, moving from rather fixed pixels and aesthetics to fluid, moving grids. These transformations gradually grew from recognizing that a diversity of devices demanded more than a single one-size-fits-all design. Today, if technological advances are starting to enable interface design (Figure 29) to adapt to individual mental models and capabilities, should not we gaze toward this possibility as actively as device compatibility was once pursued?

As we explored through the discussion of **Schramm's** model of communication, cognitive accessibility requires overlapping **fields of experience**. Some of the design approaches to accessibility primarily address perceptual barriers such as sufficient colour contrast, scalable fonts, alt-text to images or general compatibility with screen-readers and keyboard alternatives. These are absolutely necessary. But can it be assumed that once information becomes perceptible, it will become comprehensible?

Weaver's three levels of communication problems illustrate how **perceptual accessibility** (level A, how accurately can the symbols of communication be transmitted, the technical problem) does not equal **cognitive accessibility** (level B, how precisely do the transmitted symbols convey the desired meaning, the semantic problem) and its effect (level C, how effectively does the received meaning affect conduct). Understanding goes beyond perceptually receiving, although it is a pre-requisite. For many users, refining their conceptual model about the CA and engaging with the content might not only benefit but actually require customization. Learning strongly benefits from connecting to existing knowledge structures and familiar ways of thinking, and this, which is commonly used in teaching, together with what is recognized by Schramm's and Shannon-Weaver's models of communication,

is what fuels the following approach.

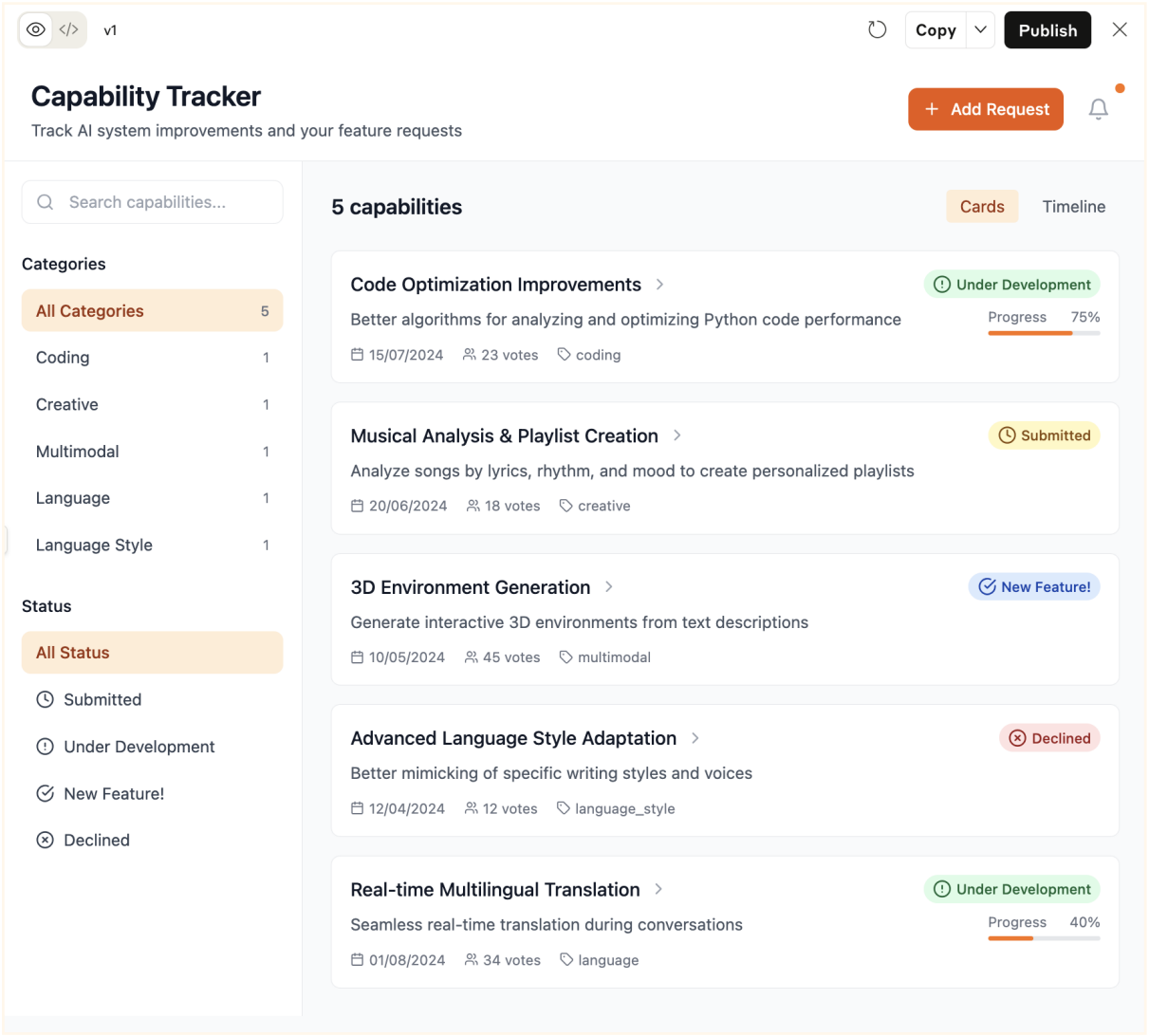


Figure 29. Sample of a first version (one-shot) design by Claude.ai having as an only briefing a very initial draft of the last paragraphs of the later section. Although there is much that we would graphically and conceptually redesign in this proposal, it represents a case in point of the potential and flexibility of these tools to quickly iterate design proposals, even with underspecified prompts. The iterative nature of CA-assisted design allows to rapidly iterate multiple concepts, and if the user is able to guide and refine the direction through conversational feedback, to choose the preferred interface elements and signifiers or request specific modifications to better align with their capabilities and desired requirements. Design generated with AI.

Approaching interfaces and their semantic content as something that can adapt also builds on our earlier analysis of the **evolution** of anthropomorphic signifiers from explicit avatars to more implicit interface layout mimicry. Just as modern CAs moved

away from human-like appearances toward borrowing familiar patterns from search engines and messaging apps, conceptual adaptation could **extend this principle** to match and assist not just with perceptual conventions but life-long experiences. The ramifications of discoverability and accessibility are a **bottleneck** on anything that follows, and thus take precedence even over those of usability and efficiency. If interfaces remain cognitively inaccessible to some people, they risk perpetrating forms of digital exclusion as CAs become more central to the daily life of many.

- Imagine a wise and inquisitive **elderly gardener** just starting to use a CA, and approaching a new AI capability report: conventional presentations assume relative familiarity with software development concepts and state-of-art AI benchmarks. Although these accurate technical terms might mean little to nothing to this persona, the gardener has decades of experience cultivating plants, understanding growth cycles, recognizing when something needs attention and caring for growth. An adaptive interface could translate these same CA capabilities into gardening metaphors for the cognitive changelog. Capabilities could become invented plants requiring different types of care, patience and tenacity, technical limitations could appear as environmental or seasonal constraints, uncertain mechanisms and interpretability challenges could be reflected by roots, and system updates could resemble fertilization or pruning, making the relevant capability development mirror the cultivation of a garden in a customized way (Figure 30, left).

- Similarly, an **expert mechanic** with extensive hands-on experience but limited access to formal technical education might benefit more from mechanical metaphors than default statistics. CA development and maintenance could take the language of automotive care and complex CA capabilities could appear mimicking engines with visible hierarchies, some components powering others and having bottlenecks visualized through flow visualization (Figure 30, right).

As we saw in the discussion section from system image to conceptual models, these design adaptations would not be merely aesthetic choices. The interface becomes a part of the system image, a message and a pedagogical tool in itself. Someone just learning mathematical notation might benefit from game-like metaphors. An elderly person intimidated by technical terminology might prefer familiar analogies from their professional or personal experience, as the gardener. Someone with **dyslexia** might benefit from spatial rather than text-heavy representations, and simplified

language structures, icons, and visual cues in that support word retrieval in the interface. Someone who cannot rely on mental imagery (such as with **aphantasia**) might opt for descriptions that build using semantic and logical constraints, or metaphorical representations systematically and explicitly shown in the interface. Similarly, someone with **blindness** engages more effectively with VUIs, but for example might also require further adaptation to audio-spatial metaphors that allow to recreate and imagine, rather than a visual card-based layout for our cognitive changelog, a soundscape where different capabilities associate with distinct tones and rhythms. Adaptive interfaces could assist diverse users from the same underlying CA, **translating and locating** what is to be communicated to make it more **cognitively accessible**.

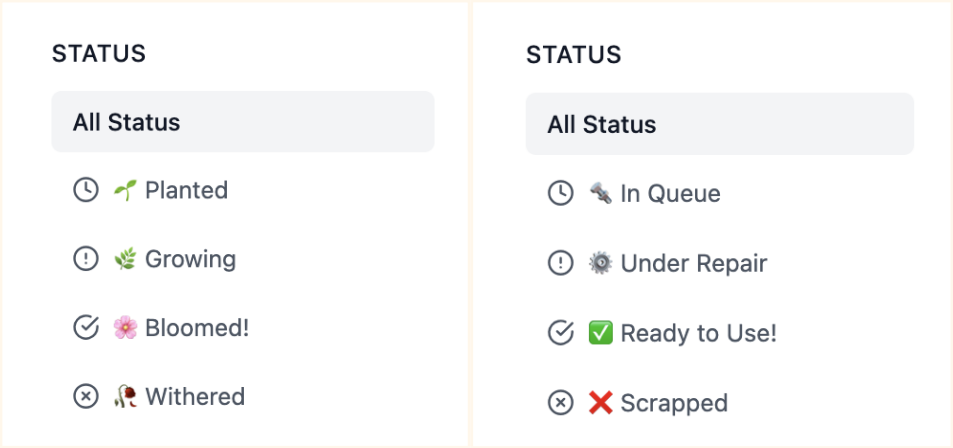


Figure 30. Fragment of a first adaptation (one-shot) for "an elderly gardener" and an "experienced mechanic", by Claude.ai. Although we do not see yet the degree of adaptation and localization that we encourage and propose, they show a starting point for optimizing and translating the more formal terms of software development to the language of different occupations. Design and labels generated with AI.

On the other side, is such a deep adaptation even technically possible and psychologically healthy? Different metaphors would function as **pedagogical tools** for accessibility, but they would also support entirely different conceptual models of how CAs operate. In our imaginary use-case personas, the garden could cue to see CA capabilities as something to cultivate, and an engine might evoke the need to tune and troubleshoot limitations. Each metaphor can be cognitive useful for understanding the intricacies of CAs and the invisible orchestra that aligns with existing expertise and fields of experience rather than requiring to build and learn entirely new conceptual models. However, approaching interface adaptation is not free of risks. Metaphors can reinforce **misleading expectations** about CAs or encourage inap-

propriate interactions, and the CA could **overadapt**, similarly to how it happened already with the verbal sycophancy problem (detailed in the invisible orchestra). Simplification can potentially obscure as much as it illuminates, be it on the side of the human or on the side of the artificial. Before implementing such transformations, extensive phases of testing would be needed to ensure that the benefits outweigh the detriments, that unintended adverse effects are accounted for, and adaptation enhances user understanding and accessibility in the desired way, as well as that the societal context is right for the sought impact.

Even more necessary would be that CAs present **multiple interface options**, allowing **users accept, modify, or dismiss** them, thus closing the **feedback** loop. This aims to mitigate the risk that adaptation is imposed rather than desired and negotiated. Like camera modes with easy automatic and complicated manual modes, the adaptive interfaces we propose could also be disabled and would offer user control over the degrees and kinds of adaptation at any moment, aiming to avoid design paternalism and the assumption that artificial systems know what users need better than they do. Customization could be suggested by the CA, but also initiated or modified by the user with natural language or examples.

At the **technical level**, Figure 29 shows a sample interface designed in one shot by Claude.ai from a rather lacking prompt, demonstrating that contemporary CAs are quite capable of generating and **iterating** visual interface design proposals **rapidly** from a minimal prompt, and Figure 30 shows how the same system can adapt textual content in the design with metaphors. If the required power capacity allows, this eventually enables the kind of customization and personal choice proposed.

Current AI systems like DeepMind's Genie can generate navigable 3D environments from descriptions in real time (24 frames per second), suggesting even more radical and enticing possibilities for expanding the fields of experience we can share with artificial systems. Future interfaces might not need to be constrained by 2D grids. Capability dashboards could become, if desired, **immersive environments** where users learn, check and explore through familiar and pedagogical spatial metaphors such as walking through a garden of imaginary plants, or examining an invented engine to understand the dependencies of the system.

Technical feasibility opens new further science fiction questions and research avenues (whose in depth exploration escapes this thesis) about interfaces and interaction design principles. When AI systems begin to evolve their own signifiers based

on user interaction patterns, this could lead to a discussion of co-evolution between human understanding and AI communication, where signifiers emerge organically rather than being deliberately designed. Furthermore, if AI would create its own interfaces, what signifiers would it choose to represent its capabilities to humans in these GUIs or mini-worlds? How might these differ from human-designed signifiers?

If the medium itself can become as fluid and adaptable as the content it presents, are we not compelled to rethink what consistency, discoverability, and agency mean in human-AI interaction? Imagine an AI-driven argumentation represented literally walking through logical paths rendered as architectural spaces, or understanding emergent capabilities by observing and interacting with them as they grow as eco-logical phenomena in simulated environments. Furthermore, an artificial system could generate entirely new experiential metaphors that do not exist in the physical world, and expand the repertoire of how humans can think about complex things, solve problems or explore questions. This kind of shift would not be unprecedented, as the Internet already altered the ways we access, process, and think about information, creating what could be called a cognitive symbiosis. Adaptive AI interfaces could extend that further.

⚑ If signifiers can be dynamic, can the entire interface be adaptive?

Perceptual accessibility (Level A, can you perceive it?) does not guarantee **cognitive accessibility** (Level B, can you make sense of it?).

- Responsive design adapts layout to devices.

→ Adaptive interfaces would adapt concepts, metaphors, and visual language to match diverse cognitive profiles and fields of experience.

- Example (signification of CA capabilities): translatable into gardening metaphors for one user, mechanical metaphors for another, soundscapes for someone with blindness, spatial representations for someone with dyslexia.
- Risks: overadaptation, loss of control, sycophancy, metaphors can reinforce misleading expectations → Adaptation negotiated, not imposed.

This page is
intentionally left blank.

6. Conclusions

This thesis has examined the graphic user interfaces (GUIs) of conversational agents (CAs) as communicative and designable artifacts, applying communication theory, cognitive science, and design. This analysis requires disciplines that rarely meet, despite having much to offer each other. What follows synthesizes what we now understand differently, and builds to what we propose as a result.

1. Communication models diagnose where meaning gets lost in human-AI interaction.

Shannon-Weaver's and Schramm's work applied to human-CA communication allows us to systematically identify "where" meaning is lost and transformed, or feedback missing, rather than simply identifying "that it is" missing.

The three levels of Weaver's communication problems (technical, semantic, and effectiveness) can be used to categorize failure modes applicable to whether signals arrive intact, whether the received signs convey the intended meaning, and whether that meaning produces the desired effect. To these three, we propose adding a fourth element: conceptual model noise. This is the distortion that accumulates when the mental models people have of the system diverge from how it actually works. A user who assumes the CA thinks, cares or reads like we do operates with a conceptual model that shapes every interpretation of every response. If it diverges too much from how the system actually functions, even technically reliable, semantically clear, effective communication can be misleading.

Communication problems in human-AI interaction occur at multiple nested levels simultaneously, combined in ways that a diagnosis at a single level misses. These distinctions have implications because their interventions differ. Technical problems require engineering solutions, but semantic and effectiveness problems require mutual and contextual understanding. A linguistically perfect reply might be a poor answer taking into account specific needs or the context of the user. Accurate but unhelpfully long responses that get only half-read might fail at effectiveness too, and prompts without context or with references to relative time or non-existent markers (such as pages in a word processor) can be misinterpreted and misuse computational resources to execute tasks in a more problematic than helpful way, which brings us to the next point.

2. Communication depends on overlapping fields of experience, and awareness.

A deceptively simple insight of Schramm's model is that effective communication requires common ground. Its etymology ("communis", shared, common) reminds us that to communicate is literally to make common, to not only find or mimic, but build shared understanding. For human-CA interaction, this implies that both parties must develop adequate models of each other. Human users need accessible conceptual models of AI capabilities, limitations and uncertainty. CAs require useful approximations of human cognition, needs, preferences, intentions and contexts. Thus, research, development, design and good use of the potential of CAs require balancing insights from cognitive science (as a discipline specifically concerned with how cognition, natural and artificial, recognizes, learns and interacts) and interpretability research (aiming to provide causal explanations of AI-generated content and the mechanisms enabling it).

The adaptation of the Johari Window to human-CA communication developed in this thesis methodically maps the asymmetry of awareness across four quadrants. Our analysis distinguishes mutually recognized limitations (Q1), blind spots on either side (Q2-Q3), and most consequentially, the mutual unknowns quadrant of what neither users nor the CA recognize (Q4, in which black swan events originate with unexpected value and unexpected harm). This highlights why interpretability research is not a choice but a requirement, as it pushes the boundaries of the unknown quadrant, potentially preventing harm that neither users nor developers could otherwise anticipate, and also recognizing new opportunities. In a way, explainable AI (XAI) efforts can be understood as attempts to expand Schramm's overlap and push the "known" Johari window quadrants, by making artificial processes more transparent and comprehensible to humans, and building understanding on what we currently cannot see.

3. Closing feedback loops and updating conceptual models through an active approach.

Feedback is necessary to close the loop that provides details to self-regulate and update our conceptual models, with a circularity that both Schramm (in communication) and Norman (in interaction design) identify. Two aggravating factors characterize the situation with CAs. First, behind the minimalistic appearance of a CA, multiple large language models (LLMs) and components of different sizes and capabilities switch depending on computational load, query type, subscription tier and updates, forming what we have called the "invisible orchestra". These changes might expand

modalities (such as readable formats), functionality (such as the context window), or bring new emerging capabilities. However, what average users see is the system image. The system image comprises interface design, documentation, the CA's self-descriptions, and accumulated user experience, which has not received investment and attention proportional to model improvements. Norman's interaction design principles show that if signifiers, mappings and constraints do not make affordances discoverable and create feedback about what actions are available and if they had the intended effect, users lack the information necessary to revise their models.

Second, if the rhythm of technological evolution is faster than the pace at which users update their beliefs of CAs, this results in a mismatched conceptual model, which echoes an Asimovian scenario (gathering knowledge faster than wisdom). The faster these technologies evolve, the wider this difference becomes. This has consequences beyond imagination and frustration. Without taking a more active stance on updating users' understanding, there is a risk that users develop trust (or distrust) based on outdated conceptual models, under-utilize new capabilities or over-rely on obsolete ones, while the invisible orchestra behind the system image transforms in ways they cannot relate to. Moreover, without indicators of gradual evolution, society may perceive AI-related transformations as discrete and abrupt and offer more resistance to what was actually incremental innovation.

The underlying opportunity, though, is that a CA, being a tool capable of flexibility, can communicate its updates and affordances, if relevant and desirable, in a customized way. This is what motivates the concept of dynamic signifiers developed later in the thesis: with AI, the interfaces of CAs can be designed to evolve with the system and the user, restoring the feedback necessary for conceptual models to stay grounded.

4. Interface hybridization creates expectations that can explain interaction patterns.

When chatbots were novel, they relied on anthropomorphic signifiers (such as human avatars or personas and simulated personality) to hint to users what kind of interaction was expected. Modern CAs evolved hybridizing the familiar aesthetics and layouts of search engines (a central text box for query, surrounded by white space, and a minimal greeting encouraging to ask anything in their homepage) and social messaging applications (typing indicators that pulse while the CA generates, chat bubbles for user messages that stack in conversation threads). Consistently with the transformation of CAs, search aesthetics called for their use as tools for information

retrieval, and as they became conversational partners for iterative work, messaging conventions made extended exchanges feel more natural.

However, borrowing signifiers entails inheriting their expectations and assumptions, and many users still implicitly expect CA responses to have the reliability of search results or believe signifiers such as the typing indicator mean what they do in human interaction, even when expectations do not extrapolate.

For future interfaces, videogame design offers an instructive example of complex and evolving systems learnt through progressive disclosure, with tutorials or interface elements that unlock as players demonstrate readiness and contextual hints that appear when relevant. Games assume their systems must be learned and design accordingly. Conversely, CA interfaces often assume they are already understood, or that understanding will happen naturally from use. The mimicry of familiar interfaces reinforces this assumption, because if it looks like searching or messaging, users will assume they know what to do.

5. Four challenges of signification.

This thesis identifies and articulates what we have termed the quadruple challenge of AI signification: the interrelated problems of signifying capabilities, limitations, and uncertainty in CAs, all entangled with the fourth, transversal challenge of interpretation. Each of these maps to how humans learn to manipulate artifacts and navigate environments. We identify affordances, recognize constraints, recalibrate on what is uncertain through feedback and create meaning across diverse contexts. These same operations map to what is most pressing to signify from a human-centered design approach (what can this do? what can this not do? how does it respond and how reliable is it? how to make sense of it?).

By grounding this in both communication models (Shannon-Weaver, Schramm) and interaction design principles (Norman), this thesis argues that these challenges are communication problems that require interdisciplinary diagnosis before engineering implementation. This analysis offers a different level of diagnosis targeted to CAs, to identify where and why communication between artificial and human cognition fails, and distinguishes between breakdowns that stem from invisible capabilities, undisclosed limitations, masked uncertainty, or interpretive divergence rooted in cultural and individual difference.

6. Dynamic signifiers and three concrete design proposals.

For systems with the described characteristics (evolving capabilities, invisible changes, context-dependent limitations, uncertainty in their mechanisms etc.) we propose the concept of dynamic signifiers. These are interface elements that evolve with the system and the interaction to communicate changing states and capabilities, actively participating in the user's evolving mental model of CAs and the discovery of affordances.

Beyond its theoretical contribution, we develop three design proposals that exemplify how dynamic signifiers might be implemented in current CA interfaces: (1) traffic light indicators that provide **contextual warning feedback** without interrupting conversational flow, (2) a **capability tracking wish list** that transforms moments of system failure into documented opportunities for user engagement and development feedback, and (3) a **cognitive changelog** that makes system evolution transparent and personally relevant through an adaptive dashboard.

They are designed with the challenges of signification in mind, including the challenge of interpretation (i.e. the changelog proposes adaptive metaphors) to be internally coherent, and they are visualized through modifying the real interfaces of current CAs (such as Claude and ChatGPT) to be coherent with external feasibility. These proposals are offered as preliminary explorations rather than validated solutions, each requiring empirical testing with diverse user populations. Nevertheless, they demonstrate that the signification challenges identified in this thesis are not merely theoretical but can be put into practice with specific, implementable design directions.

⚑ Summary: Interface design for AI is a communication problem

The signification challenges identified in this thesis (signifying capabilities, limitations, uncertainty, and navigating interpretation) are communication problems that require interdisciplinary diagnosis before they require engineering implementation.

Communication models (Shannon-Weaver, Schramm) diagnose where meaning is lost, not just that it is. Interaction design principles (Norman) show how to make what is invisible discoverable.

Together, they reveal that:

- If the rhythm of technological evolution is faster than the pace at which users update their conceptual models, the difference between gathering wisdom and generated knowledge widens (Asimov).
- The system image of CAs has not received investment proportional to model improvements. Small interface changes (a clearer icon, a warning indicator, adaptive metaphors) can transform millions of daily interactions.
- Interpretability research is not optional. It expands Schramm's overlap and pushes the Johari Window's unknown quadrants toward the known, preventing harm that neither users nor developers could otherwise anticipate.

The interfaces of CAs are communicative artifacts. Designing them is a communicative act, and currently, an underused one.

7. Limitations and Further Research

Everything presented in this thesis is a contribution to a broader field of research on artificial cognition, human understanding, and the interfaces that mediate between them, and with the environments in which they move. This research needs two complementary directions: that of **what we do not see** (the internal mechanisms of AI, whose opacity and complexity create the explainability paradox) and that of **what we see** (the system image: the interfaces, representations, and languages, verbal and visual, through which these systems present themselves to users).

Researching the invisible orchestra (internal mechanisms) is necessary because interpretability is needed before explainability can be put into practice (in order to explain a process, it is necessary to understand it first). This is another important limitation, as the **side of what we cannot see** remains beyond the scope of this thesis, and is being actively researched [52][59]. Recent work on mechanistic interpretability [60], inner interpretability informed by **cognitive neuroscience** [61], internal representations [62], and multilevel analysis of ANNs suggest that progress is possible [63], and that the conceptual tools and methods of cognitive science, such as **Marr's three levels of analysis** applied by He et al. to **AI interpretability**, can help organize this effort and identify research gaps.

Researching the system image is necessary because complete expert understanding alone does not ensure that knowledge will serve the people who need it. Thus, this thesis focuses on the interactive side of **what we can see**: particularly the GUI as a communicative artifact. Other elements of the **system image**, like external documentation and tutorials on CAs, or the websites of CA companies, are beyond the scope of our work but require careful attention, as they directly inform our conceptual models too. Broader interfaces, including embodied **human-robot interactions**, are unaddressed too, even if interaction design principles, anthropomorphism or the hybridization of interfaces find informative parallels in human-robot interaction, and further integrative research would be beneficial. At a theoretical level, **transactional models** such as Barnlund's could complement the analysis of communication, and **semiotic** theory could provide a more systematic account of signs. At the level of implementation, user experience research across different cultures, populations, and levels of expertise would be needed to validate and **prototype the dynamic signifiers and design proposals** presented in chapter 5, as they have not been empirically tested and refined yet.

Another point to consider is that, as AI keeps progressing, it seems to acquire a wider field of experience and knowledge, while human limitations remain rather invariant. **Accessibility guidelines** and **heuristics** specifically targeted to CA design may need to be developed for inclusive and adaptable design. Existing web accessibility principles might not transfer directly to CAs, as turn-taking or the simulation of agency and introspection introduce novel dimensions absent from GUI-based standards [64][65].

Finally, the analysis of interface design has centred on Norman's principles of interaction design, but the natural continuation of this thesis work is a **multi-level analysis of the visual language for interface design**, analogous to the levels of analysis in linguistics. This comprises, in addition to Norman's principles (how meaning is constructed through interaction), **Nielsen's usability heuristics** (practical guidelines grounded in cognitive patterns) and interface conventions, **Gestalt principles of visual perception** (how elements are perceived as organised patterns, akin to syntax), and the **atomic elements of visual language** (point, line, direction, shape, texture, and color; the irreducible units of visual communication akin to phonemes and morphemes). Such an approach builds on the premise of GUIs as communication artifacts, but deepens by providing the pieces to design more consciously, with awareness of the vocabulary, grammar, and contextual considerations of visual language. Within the field of cognitive science, this approach draws on research about visual thinking. Arnheim's work and related thought experiments ask, for example, whether solving spatial mental rotation requires visual imagination or abstract reasoning, or if the two are constructs of an inseparable skill [66].

We are confident that cognitive science, in collaboration with other disciplines, such as design and engineering, will benefit from these further research directions significantly. At a societal level, the findings of these research initiatives will be important assets for us as humans who are increasingly more embedded in artificial environments.

References

[0] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. USA: Henry Holt and Co., Inc., 1982.

[1] "ISO 9241-210:2019, Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems." International Organization for Standardization, 2019. Available: *https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en*. [Accessed: Apr. 05, 2025]

[2] Wikipedians, "User interface," Wikipedia. Mar. 19, 2025. Available: *https://en.wikipedia.org/w/index.php?title=User_interface*. [Accessed: Apr. 05, 2025]

[3] M. Worthington, "What is a User Interface?," *California Institute of the Arts. in Visual Elements of User Interface Design. Available: https://www.coursera.org/learn/visual-elements-user-interface-design/home/module/1*. [Accessed: Apr. 06, 2025]

[4] "Model Context Protocol (MCP)," *Anthropic. Available: https://docs.anthropic.com/en/docs/agents-and-tools/mcp*. [Accessed: Apr. 06, 2025]

[5] J. R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences, vol. 3, no. 3, pp. 417–424, 1980, doi: 10.1017/S0140525X00005756*

[6] D. Silver, T. Hubert, J. Schrittwieser, D. Hassabis, "AlphaZero: Shedding new light on chess, shogi, and Go," Google DeepMind, Dec. 06, 2018. Available: *https://deepmind.google/discover/blog/alphazero-shedding-new-light-on-chess-shogi-and-go/*. [Accessed: Feb. 16, 2025]

[7] AlphaStar team, "AlphaStar: Grandmaster level in StarCraft II using multi-agent reinforcement learning," Oct. 30, 2019. Available: *https://deepmind.google/discover/blog/alphastar-grandmaster-level-in-starcraft-ii-using-multi-agent-reinforcement-learning/*. [Accessed: Feb. 16, 2025]

[8] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, D. Silver, "MuZero: Mastering Go, chess, shogi and Atari without rules," Dec. 23, 2020. Available: *https://deepmind.google/discover/blog/muzero-mastering-go-chess-shogi-and-atari-without-rules/*. [Accessed: Feb. 16, 2025]

[9] D. Hassabis, "Future of AI, Simulating Reality, Physics and Video Games," Jul. 23, 2025. Available: *https://www.youtube.com/watch?v=-HzgcbRXUK8*. [Accessed: Aug. 16, 2025]

[10] Z. M. Beddoes, *Interview With Demis Hassabis and Dario Amodei: AI Bosses on What Keeps Them up at Night,* (Feb. 18, 2025). Available*: https://www.youtube.com/watch?v=4poqjZlM8Lo*. [Accessed: Aug. 16, 2025]

[11] D. C. Dennett, *The Intentional Stance*. MIT Press and Bradford Books, 1989.

Available: https://mitpress.mit.edu/9780262540537/the-intentional-stance/. [Accessed: Apr. 06, 2025]

[12] D. Harper and T. Felix, "Origin and history of artificial," *Online Etymology Dictionary (Etymonline)*. Available: https://www.etymonline.com/word/artificial. [Accessed: Apr. 14, 2025]

[13] C. Shannon and W. Weaver, "The Mathematical Theory of Communication," University of Illinois Press, 1949.

[14] C. Shannon, "A Mathematical Theory of Communication," 1948.

[15] A. G. Basile and M. E. Toplak, "Four converging measures of temporal discounting and their relationships with intelligence, executive functions, thinking dispositions, and behavioral outcomes," Front. Psychol., vol. 6, Jun. 2015, doi: 10.3389/fpsyg.2015.00728. Available: http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.00728/abstract. [Accessed: Apr. 11, 2025]

[16] L. Steels, "Evolving grounded communication for robots," *Trends in Cognitive Sciences, vol. 7, no. 7, pp. 308–312, Jul. 2003, doi: 10.1016/S1364-6613(03)00129-3*. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661303001293. [Accessed: Feb. 15, 2026]

[17] L. Steels, *The Talking Heads experiment: Origins of words and meanings*. Language Science Press, 2015. doi: 10.26530/OAPEN_559870. Available: https://library.oapen.org/handle/20.500.12657/33193. [Accessed: Feb. 15, 2026]

[18] W. Schramm, "Procedures and Effects of Mass Communication," Teachers College Record: The Voice of Scholarship in Education, vol. 55, no. 10, pp. 113–138, Aug. 1954, doi: 10.1177/016146815405501006. Available: https://journals.sagepub.com/doi/10.1177/016146815405501006. [Accessed: Apr. 05, 2025]

[19] University of Maryland, Wisconsin Historical Society, Maryland Institute for Technology in the Humanities, "Biography of Wilbur Schramm," *Unlocking the Airwaves. Available:* https://www.unlockingtheairwaves.org. [Accessed: Apr. 13, 2025]

[20] E. W. Schramm and D. F. Roberts, "The Process and Effects of Mass Communication (Revised Edition)," *University of Illinois Press, 1971, doi: 10.1177/002194367200900310*. Available: https://www.worldradiohistory.com/BOOKSHELF-ARH/Education/The-Process-and-Effects-of-Mass-Communications-Schramm-1971.pdf

[21] J. von Uexküll, *Umwelt und Innenwelt der Tiere*. Berlin : J. Springer, 1909. Available: http://archive.org/details/umweltundinnenwe00uexk. [Accessed: Apr. 13, 2025]

[22] T. Nagel, "What Is It Like to Be a Bat?," *The Philosophical Review, vol. 83, no. 4, p. 435, Oct. 1974, doi: 10.2307/2183914*. Available: https://www.jstor.org/stable/2183914?origin=crossref. [Accessed: Apr. 14, 2025]

[23] "How I use sonar to navigate the world," *TED2015, TED Talks, Mar. 2015. Available: https://www.ted.com/talks/daniel_kish_how_i_use_sonar_to_navigate_the_world*. [Accessed: Apr. 14, 2025]

[24] J. C. Maxwell, "On Governors," Proceedings of the Royal Society, vol. XVI, no. 100, 1868.

[25] N. Wiener, *Cybernetics or Control and Communication in the Animal and the Machine*. The MIT Press, 1961. doi: *10.7551/mitpress/11810.001.0001*. Available: *https://direct.mit.edu/books/oa-monograph/4581/Cybernetics-or-Control-and-Communication-in-the*. [Accessed: Apr. 16, 2025]

[26] J. von Neumann, "The Mathematician," *Works of the Mind, vol. 1, no. 1, 1947, Available: https://math.bme.hu/~petz/vnmathematician.html*. [Accessed: Feb. 16, 2026]

[27] H. A. Simon, *The sciences of the artificial, 3. ed., [Nachdr.]. Cambridge, Mass.: MIT Press, 2008.*

[28] K. Krippendorff, *The Semantic Turn: A New Foundation for Design*. Routledge (CRC Press), 2006. Available: *https://www.routledge.com/The-Semantic-Turn-A-New-Foundation-for-Design/Krippendorff/p/book/9780415322201*. [Accessed: Feb. 15, 2026]

[29] R. Buchanan, "Wicked Problems in Design Thinking," *Design Issues, vol. 8, no. 2, p. 5, Spring 1992, doi: 10.2307/1511637*. Available: *https://www.jstor.org/stable/1511637?origin=crossref*. [Accessed: Feb. 15, 2026]

[30] H. W. J. Rittel and M. M. Webber, "Dilemmas in a General Theory of Planning," *Policy Sciences, vol. 4, no. 2, pp. 155–169, 1973, Available: http://www.jstor.org/stable/4531523*

[31] D. A. Norman, *The design of everyday things, Rev. and Expanded edition. Cambridge (Mass.): MIT press, 2013.*

[32] D. A. Norman, "Affordance, conventions, and design," interactions, vol. 6, no. 3, pp. 38–43, May 1999, doi: *10.1145/301153.301168*. Available: *https://dl.acm.org/doi/10.1145/301153.301168*. [Accessed: Apr. 21, 2025]

[33] D. A. Norman, "Signifiers, not affordances," *interactions, vol. 15, no. 6, pp. 18–19, Nov. 2008, doi: 10.1145/1409040.1409044*. Available: *https://dl.acm.org/doi/10.1145/1409040.1409044*. [Accessed: Apr. 29, 2025]

[34] D. A. Norman and S. W. Draper, Eds., "User Centered System Design: New Perspectives on Human-computer Interaction," *Routledge & CRC Press, 1986, Available: https://www.routledge.com/User-Centered-System-Design-New-Perspectives-on-Human-computer-Interaction/Norman-Draper/p/book/9780898598728*. [Accessed: Apr. 30, 2025]

[35] M. Chiriatti, M. Ganapini, E. Panai, M. Ubiali, and G. Riva, "The case for human–AI interaction as system 0 thinking," Nat Hum Behav, vol. 8, no. 10,

pp. 1829–1830, Oct. 2024, doi: 10.1038/s41562-024-01995-5. Available: https://www.nature.com/articles/s41562-024-01995-5. [Accessed: Apr. 05, 2025]

[36] "Don Norman Design Award (DNDA) and Summit." Available: https://dnda.design/about/#behind-dnda. [Accessed: Apr. 30, 2025]

[37] J. J. Gibson, "The Senses Considered as Perceptual Systems," *The Quarterly Review of Biology, vol. 44, no. 1, pp. 104–105, Mar. 1969, doi: 10.1086/406033*. Available: https://www.journals.uchicago.edu/doi/10.1086/406033. [Accessed: Apr. 29, 2025]

[38] J. J. Gibson, *The Ecological Approach to Visual Perception,* Psychology Press. Taylor & Francis, 1986.

[39] W. W. Gaver, "Technology affordances," in Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91, New Orleans, Louisiana, United States: ACM Press, 1991, pp. 79–84. doi: 10.1145/108844.108856. Available: http://portal.acm.org/citation.cfm?doid=108844.108856. [Accessed: Apr. 29, 2025]

[40] R. Hartson, "Cognitive, physical, sensory, and functional affordances in interaction design," Behaviour & Information Technology, vol. 22, no. 5, pp. 315–338, Sep. 2003, doi: 10.1080/01449290310001592587. Available: http://www.tandfonline.com/doi/abs/10.1080/01449290310001592587. [Accessed: Feb. 18, 2026]

[41] J. McGrenere and W. Ho, "Affordances: Clarifying and Evolving a Concept," in Proceedings of Graphics Interface 2000, Montréal, Québec, Canada, May 2000, pp. 179–186.

[42] R. E. Núñez and E. Sweetser, "With the Future Behind Them: Convergent Evidence From Aymara Language and Gesture in the Crosslinguistic Comparison of Spatial Construals of Time," Cognitive Science, vol. 30, no. 3, pp. 401–450, 2006, doi: 10.1207/s15516709cog0000_62. Available: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0000_62. [Accessed: Feb. 15, 2026]

[43] L. Boroditsky and A. Gaby, "Remembrances of Times East: Absolute Spatial Representations of Time in an Australian Aboriginal Community," *Psychol Sci, vol. 21, no. 11, pp. 1635–1639, Nov. 2010, doi: 10.1177/0956797610386621*. Available: https://doi.org/10.1177/0956797610386621. [Accessed: Feb. 15, 2026]

[44] J. Weizenbaum, "Contextual Understanding by Computers," *Commun. ACM, vol. 10, no. 8, pp. 474–480, Aug. 1967, doi: 10.1145/363534.363545*. Available: https://dl.acm.org/doi/10.1145/363534.363545. [Accessed: Jun. 14, 2025]

[45] K. M. Colby, F. D. Hilf, S. Weber, and H. C. Kraemer, "Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes," *Artificial Intelligence, vol. 3, pp. 199–221, Jan. 1972, doi: 10.1016/0004-3702(72)90049-5*. Available: https://www.sciencedirect.com/science/article/

pii/0004370272900495. [Accessed: Jun. 14, 2025]

[46] J. Luft, *Of human interaction*. Palo Alto, Calif. : National Press Books, 1969. Available: *http://archive.org/details/ofhumaninteracti00jose*. [Accessed: Aug. 20, 2025]

[47] L. C. Porter and B. Mohr, *Reading Book for Human Relations Training*. NTL Institute, 1982.

[48] A. L. Cooke, M. Brazzel, A. S. Craig, and B. Greig, Eds., *Reading Book for Human Relations Training*, Eighth Edition. Alexandria, Va: NTL Institute, 1999.

[49] M. Krogerus and R. Tschäppeler, *The Decision Book: 50 Models for Strategic Thinking*. Erscheinungsort nicht ermittelbar: W. W. Norton & Company, 2012.

[50] "Department of State Washington File Transcript: Defense Department Briefing, February 12." Pentagon, Dec. 02, 2002. Available: *https://usinfo.org/wf-archive/2002/020212/epf202.htm*. [Accessed: Aug. 20, 2025]

[51] M. Shanahan and J. Bi, "Human and Machine Consciousness," Apr. 2025. Available: *https://www.youtube.com/watch?v=bBdE7ojaN9k* [Accessed: May 14, 2025]

[52] J. Lindsey *et al.*, *"On the Biology of a Large Language Model,"* Transformer Circuits Thread, Mar. 2025, Available: *https://transformer-circuits.pub/2025/attribution-graphs/biology.html, https://www.anthropic.com/research/tracing-thoughts-language-model*. [Accessed: Jul. 30, 2025]

[53] R. Greenblatt *et al.*, *"Alignment faking in large language models." arXiv, Dec. 20, 2024. doi: 10.48550/arXiv.2412.14093*. Available: *http://arxiv.org/abs/2412.14093, https://www.anthropic.com/research/alignment-faking*. [Accessed: Jul. 30, 2025]

[54] Participants of the Anthropic Fellows program, "Persona vectors: Monitoring and controlling character traits in language models," Jan. 08, 2025. Available: *https://www.anthropic.com/research/persona-vectors*. [Accessed: Aug. 17, 2025]

[55] R. Chen, A. Arditi, H. Sleight, O. Evans, and J. Lindsey, "Persona Vectors: Monitoring and Controlling Character Traits in Language Models." arXiv, Jul. 29, 2025. doi: *10.48550/arXiv.2507.21509*. Available: *http://arxiv.org/abs/2507.21509*. [Accessed: Aug. 17, 2025]

[56] W. W. A. Initiative (WAI), "WCAG 2 Overview." Available: *https://www.w3.org/WAI/standards-guidelines/wcag/*. [Accessed: Aug. 23, 2025]

[57] Bureau of Internet Accessibility, "Does WCAG 2.1 Require Responsive Design?" Aug. 18, 2022. Available: *https://www.boia.org/blog/does-wcag-2.1-require-responsive-design*. [Accessed: Aug. 23, 2025]

[58] "HTML Responsive Web Design," *W3Schools. Available: https://www.*

w3schools.com/html/html_responsive.asp. [Accessed: Aug. 23, 2025]

[59] J. Lindsey, "Emergent Introspective Awareness in Large Language Models," Transformer Circuits Thread, Oct. 2025, Available: https://transformer-circuits.pub/2025/introspection/index.html. [Accessed: Mar. 04, 2026]

[60] L. Bereska and E. Gavves, "Mechanistic Interpretability for AI Safety -- A Review." arXiv, Aug. 23, 2024. doi: 10.48550/arXiv.2404.14082. Available: http://arxiv.org/abs/2404.14082. [Accessed: Apr. 05, 2025]

[61] M. G. Vilas, F. Adolfi, D. Poeppel, and G. Roig, "Position: An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience." arXiv, July 31, 2024. doi: 10.48550/arXiv.2406.01352. Available: http://arxiv.org/abs/2406.01352. [Accessed: Apr. 05, 2025]

[62] W. Gurnee and M. Tegmark, "Language Models Represent Space and Time." arXiv, Mar. 04, 2024. doi: 10.48550/arXiv.2310.02207. Available: http://arxiv.org/abs/2310.02207. [Accessed: Apr. 05, 2025]

[63] Z. He et al., "Multilevel Interpretability Of Artificial Neural Networks: Leveraging Framework And Methods From Neuroscience." arXiv, Aug. 26, 2024. doi: 10.48550/arXiv.2408.12664. Available: http://arxiv.org/abs/2408.12664. [Accessed: Apr. 05, 2025]

[64] C. Murad, C. Munteanu, L. Clark, and B. R. Cowan, "Design guidelines for hands-free speech interaction," in Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, Barcelona Spain: ACM, Sept. 2018, pp. 269–276. doi: 10.1145/3236112.3236149. Available: https://dl.acm.org/doi/10.1145/3236112.3236149. [Accessed: Apr. 05, 2025]

[65] M. Skjuve, A. Følstad, and P. B. Brandtzaeg, "The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users," in Proceedings of the 5th International Conference on Conversational User Interfaces, Eindhoven Netherlands: ACM, July 2023, pp. 1–10. doi: 10.1145/3571884.3597144. Available: https://dl.acm.org/doi/10.1145/3571884.3597144. [Accessed: Apr. 05, 2025]

[66] R. Arnheim, "A Plea for Visual Thinking," Critical Inquiry, vol. 6, no. 3, pp. 489–497, Apr. 1980, doi: 10.1086/448061. Available: https://www.journals.uchicago.edu/doi/10.1086/448061. [Accessed: Mar. 04, 2026]

Acknowledgements

Acknowledgement of humans

I want to extend special thanks to:

First and foremost, my mentor Martin, for his patient guidance that persisted over the course of writing this thesis, as well as his trust in me and my work, for always being kind, and for showing me that a thesis, like a book, is a work of art and love.

Alex, for being the constant in my life. I will never thank you enough for the peace, clarity, and support you bring to our everyday.

Ekin, because without our interaction and your care, this thesis and I would not be what we are, and would not be.

My parents, for teaching me the importance of explanation, communication, and understanding, and my brother, for being my first editor, in thesis as in life.

My friends, including but not limited to: Digvijay, whose admirable ambition and methods helped me overcome a block and start collecting thoughts in what became a preliminary draft. Radu, with whom we dream and pursue everyday utopias for better and more customizable software. Marie and Udo, for our master thesis squad, and for being an inspiration with their own theses, commitment, and creative lives. Anaïs, whose tenacity and organized visualization trackers gave closure to this thesis and the last months of work. Tessa, whose puzzle and attitude are a permanent reminder of the value of innovation and exploration. Àngels and Stefan, who inspired me to meditate on ideas and life, and without whom I would not have found the space to resolve some chapters. And all others who are part of this journey, which has been its own destination, has adapted to the ever-changing hereness, increased the area of serendipity, and hopefully avoided a "failure of imagination."

Thank you. Ďakujem. Graia. Teşekkürler. Gràcies. धन्यवाद्. Mulțumesc. Danke. For being how you are, and a part of my life.

Thanks to all the people who made and continue to make MEi:CogSci what it is. Especially to Lisl, whose strength, perseverance, and timely improvisation made complex choreography a living dance. And to Markus, whose openness and clarity inspired me from day one to innovate and think further, and brought me to explore cognition integratively, as well as to discover affordances. This thesis could not have been born anywhere else but in a study program that brings cohesion and breadth

to explore neurostimulation research in one of the biggest hospitals in Europe, build a robot-controlling glove prototype soldering gyroscopes and connecting bend sensors to an Arduino board, gather data for dynamic eye-tracking studies in art museums, and write papers about optical illusions for AI, the morality of tools, how children build their own concept of love in different cultures, or the interplay between conscious experience and writing.

Last but not least, to all those anonymous people out there, like the kind employee at a Lego store who let me take more pieces than allowed for my character, which served to understand constraints (Figure 17). Sometimes research and creativity require kindness, and thoughtful exceptions.

Acknowledgement of tools

This thesis was written with the support of several software tools, each of which deserves its own quiet acknowledgement. Obsidian and its plugins, for writing it. Microsoft Word, for reviewing and commenting on parts of it. Zotero and its Firefox extension, for compiling research and citing it later. Adobe Illustrator and Photoshop, for drawing, vectorizing, and editing the figures. Adobe InDesign, for the layout and design you are reading now. And AI tools: NotebookLM, used as an advanced PDF search to find relevant content across books and support a referenced study of sources. Gemini and ChatGPT, to transcribe speech into text. And Claude, both as a conversational partner and a technology always ready to provide feedback on what I write, its synonyms, interpretation, and alternatives. These, together with others (including Mistral's Le Chat, DeepSeek, and Perplexity), also served to test and experience on myself the interaction patterns discussed in this thesis.

Thank you to the teams that made all these tools a reality of today. And those making the ones of tomorrow.

This page is
intentionally left blank.

Curious where this goes next?

www.elisabetdelgadomas.com