POLICY BRIEF

Living with AI

A Philosophical Toolkit for Navigating the Conceptual Challenges of Artificial Intelligence Systems

---

Institut Jean Nicod
Département d'Études Cognitives, École Normale Supérieure - PSL; École des hautes études en sciences sociales

Authors: Roberto CASATI and Piera MAURIZIO
Editors: Quentin COUDRAY, Alda MARI, Gloria ORIGGI

Paris, 18 December 2025

Institut | Nicod

Abstract

This policy brief examines how contemporary AI systems (especially generative systems used in everyday, non-expert interactions) destabilize core concepts through which societies organize responsibility, meaning, creativity, and agency. Rather than treating these disruptions as primarily technical or regulatory problems, the brief argues that they are fundamentally conceptual. It proposes a strategy of conceptual forcing: for the sake of practical governance, current and foreseeable AI systems are treated as non-conscious, non-creative, non-responsible, and non-intentional, regardless of their appearance. Through four case studies (consciousness, creativity, meaning, and personhood), the brief shows how AI systems generate powerful illusions of mentality that shape user expectations and social practices, often independently of explicit beliefs. These effects arise from design and training choices, not from machine cognition. The brief argues that effective AI governance therefore requires integrating conceptual engineering into design, communication, and policy, in order to preserve epistemic integrity, prevent responsibility displacement, and guide human-AI interaction in socially and democratically grounded ways.

About the Jean Nicod Institute:

A joint research unit of the CNRS founded in 2002, with École normale supérieure (ENS) and the École des hautes études en sciences sociales (EHESS) as its two academic supervising institutions, the Jean Nicod Institute is an interdisciplinary cognitive science laboratory comprising around one hundred members. Its unifying focus is the human mind and the nature of representations—linguistic, mental, and social. www.institutnicod.org.

Corresponding author: Piera Maurizio (piera.maurizio@ens.psl.eu)

This work was carried out within the framework of the Horizon Europe project ASTOUND.

To cite this policy brief: Casati, R. & Maurizio, P. (2025). Living with AI. A Philosophical Toolkit for Navigating the Conceptual Challenges of Artificial Intelligence Systems. Zenodo. https://doi.org/10.5281/zenodo.18000868

Table of contents

Executive Summary **4**

Rationale **6**

1. Introduction: A Shifting Landscape - AI and Human Concepts **8**

2. Methodology: Conceptual Engineering as a Tool for Navigating the AI-reshaped information ecosystem **10**

3. Case Studies: AI as a Limit Case for Consciousness, Creativity, Text Meaning and Personhood **12**

a. Consciousness: living with a powerful illusion 12

b. Creativity: Not just novelty 15

c. Meaning: texts vs. quasi-texts 17

d. Personhood 20

4. Policy recommendations **24**

1) Promote Conceptual Hygiene in Public Discourse 24

2) Integrate Conceptual Engineering into Policy Design 24

3) Guard Against Misleading Anhropomorphism 24

4) Reinforce the recognition of the Human Role in Creative and Communicative Acts 25

5) Monitor and Protect Epistemic Environments 25

6) Support Shared Conceptual Infrastructure 25

Bibliography **27**

Executive Summary

Living with Artificial Intelligence: Conceptual Engineering for Everyday Human–AI Interaction

Artificial intelligence systems, especially generative AI such as chatbots, text generators, and image models, are increasingly embedded in everyday life. Ordinary users now routinely interact with systems that produce fluent language, convincing images, and context-sensitive responses. These systems challenge some of our most fundamental concepts: consciousness, creativity, meaning, agency, and personhood. While technical and legal debates around AI are advancing rapidly, our shared conceptual framework for understanding and governing these systems has not kept pace.

This policy brief argues that many current difficulties in AI governance stem from conceptual confusion. We often talk about AI using categories originally developed to describe human mental life, even when those categories no longer apply straightforwardly. As a result, public discourse, design choices, and policy debates risk being driven by misleading metaphors and speculative futures rather than by the realities of how AI systems function and how people actually interact with them.

Scope and Approach

The brief focuses deliberately on **everyday interactions between non-expert users and AI systems**, such as conversational chatbots, generative text and image tools, and assistive applications used in education, communication, and cultural production. It does not address military, industrial, or highly specialized professional uses of AI, although many of the conceptual tools developed here may later be extended to those domains.

Methodologically, the brief adopts **conceptual forcing**: a pragmatic philosophical strategy that stipulates clear working assumptions in order to enable concrete reasoning and decision-making. In particular, the brief proceeds on the assumption that current AI systems are **not conscious, not creative in the human sense, not moral agents, and not bearers of meaning or responsibility** – even though they are often perceived as such by users. The central question, therefore, is not what AI systems "really are," but **how we should live with machines that convincingly simulate human-like capacities**.

Key Findings

Through four case studies – **consciousness, creativity, meaning, and personhood** – the brief shows how AI systems generate powerful illusions that shape user behavior, trust, and social expectations:

- **Consciousness:** AI systems simulate attention, memory, and emotional responsiveness, triggering intuitive attributions of sentience. These attributions are driven by surface cues and interactional design, not by genuine experience.
- **Creativity:** AI systems generate novel and valuable outputs without intention or expressive aims, destabilizing traditional criteria for authorship, originality, and artistic value.

- **Meaning:** AI-generated texts resemble human communication but lack communicative intent and truth commitment, producing "quasi-texts" that risk polluting epistemic environments.
- **Personhood:** Treating AI systems as persons can blur responsibility and displace accountability from designers and institutions to machines that cannot be morally responsible.

Across all cases, the core risk is not metaphysical error but **epistemic and normative drift**: over-trust, a-critical deference, responsibility misattribution, and erosion of practices that sustain human agency, interpretation, and judgment.

Policy Orientation

The brief argues that these challenges arise **at the point of design**, not only at deployment or use. Design choices – both in system architecture and in training data – shape how users interpret AI systems and how social norms evolve around them. Conceptual engineering must therefore be integrated upstream into AI development and governance.

Rather than proposing a comprehensive regulatory framework, the brief offers **actionable policy recommendations** organized around six priorities:

1. Promoting conceptual hygiene in public discourse
2. Integrating conceptual engineering into policy design
3. Guarding against misleading anthropomorphism
4. Reinforcing recognition of the human role in creative and communicative acts
5. Monitoring and protecting epistemic environments
6. Supporting shared, evolving conceptual infrastructure

These recommendations align with emerging legal frameworks, including the EU AI Act, while emphasizing that regulation alone is insufficient without sustained conceptual clarity.

Conclusion

Living with AI requires more than technical safeguards or compliance mechanisms. It requires **re-engineering the concepts through which we interpret, design, and govern artificial systems**. By clarifying what is at stake when we invoke notions such as consciousness, creativity, meaning, and personhood, conceptual engineering can help policymakers, designers, and users navigate AI's societal impact with greater precision, responsibility, and democratic accountability.

Living with AI

A Philosophical Toolkit for Navigating the Conceptual Challenges of Artificial Intelligence Systems

Rationale

Philosophers are currently - and insistently - asked to provide answers to questions involving Artificial Intelligence (AI, a broad notion encompassing the computational architectures behind certain types of chatbots, machine translators, text completion assistants, recommender systems, visual recognition systems, among others.) The questions involve the attribution of moral responsibility, creativity, consciousness, personhood, agency, and meaning, among others.

Does a string of characters produced by a chatbot *mean* anything? Will a machine ever be *conscious*? Who is *responsible* for a traffic casualty involving what appears to be a *decision* made by an autonomous vehicle? Do machines *see* objects; do they *recognize* faces? Do we call 'reasoning' a process that must resemble human reasoning (i.e., that includes various types of biases), and if so, how should we call a machine process that is not prone to those biases: is it a reasoner after all? We sometimes say that and AI should be considered 'at best' as an *assistant*, which kind of relationship is this?

Much as we consider these questions as potentially interesting, and certainly difficult, in this brief we propose that good theorizing and decision-making about AI requires "conceptual forcing", i.e., it requires that strong assumptions be made about the concepts used to talk about AI in order to avoid cross-talk and insoluble theoretical oppositions. Conceptual forcing thus offers a practical solution to promote concrete decision-making *now* (possibly at the expense of being assumptions that could ultimately be refuted by further technological advances). In particular, we "conceptually force" the following two steps:

- first, we take for granted a battery of bottom-line skeptical responses to the key questions (for instance, we assume that AI is not conscious now and that the issue of its potentially becoming conscious in the (far) future is irrelevant to theorizing and policymaking about it, and this in the face of our tendency to consider it conscious in certain circumstances);

- and second, we explore how to live with machines that (according to the first step) we assume to be "just machines".

The second step of the forcing is a form of conceptual *engineering*, calling for the introduction of a toolbox of concepts for negotiating the transition to a philosophically mature understanding and use of AI.

The forcing will be probably felt as such because we are prone to the tendency to take an intentional stance towards computers - a tendency that is reinforced by the design of the interfaces we use to interact with the relevant machines. In the present brief we are not delving deep into the causes of the attribution problem (e.g. the automaticity of an intentional stance towards machines, or of the advantages of using anthropomorphic language etc.). Neither are we endorsing the adage that "if you know it, you'll avoid it": a diagnosis of the problem is not a therapy. Our stance is to make concrete proposals to live

with an AI that we can safely consider as non-conscious, non-creative, irresponsible, that produces meaningless text, and to live this life *in spite of* our tendency to consider IA conscious, responsible, creative and meaningful. Some analogies will help us regulate this interaction, in particular the idea that we should treat machines like members of an alien culture that is colonizing our lives and that we are trying to understand.

Scope note

Although Artificial Intelligence is deployed across a wide range of domains – including professional decision-making, social media governance, industrial automation, and military applications – this brief deliberately focuses on **everyday interactions between non-expert users and generative AI systems**, such as chatbots and other generative interfaces. These systems are currently the primary site where conceptual confusion around consciousness, creativity, meaning and personhood arises in ordinary social contexts. The analyses and recommendations that follow should therefore be understood as addressing this specific domain of use, rather than the full spectrum of AI applications.

1. Introduction: A Shifting Landscape - AI and Human Concepts

The rapid development of Artificial Intelligence (AI), or Intelligences, particularly in the domain of Generative Artificial Intelligence, is transforming our social, economic, and geo-political environment in unprecedented ways. At a broader level, AIs are alleged to reshape our society, disrupt labor markets, challenge educational norms, alter the fabric of communication and information-sharing, and contribute to generalized brain rot. But it is at the individual level that their influence is most immediate, since these technologies are increasingly shaping our everyday experiences: from search engines that anticipate our queries, to chatbots purporting to offer emotional support, to image generators that blur the boundaries between human-made and machine-made art.

This transformation raises a wide range of complex issues (philosophical, cognitive, ethical, legal, and political) and exerts pressure on many of the core concepts we rely on to make sense of our world. Terms like 'creativity', 'meaning', 'consciousness', 'intelligence', 'agency', and 'personhood', once relatively stable, appear now to be in flux - that was of course Alan Turing's prediction:

"*The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.*"(Turing 1950)

And as AI systems increasingly perform tasks that once required human judgment or expression, we lack a shared conceptual framework to interpret these developments, let alone to design well-grounded norms capable of addressing them.

As new AI-driven phenomena emerge, we often find ourselves without the appropriate conceptual tools to assess their meaning, value, or risks. For instance, the mass production of AI-generated text, images, and sound is reshaping our informational landscape, but are we prepared to rethink notions of *authorship*, *trustworthiness*, or *originality* in light of this change? Similarly, the integration of AI into decision-support systems – in healthcare, education, criminal justice, and beyond – is felt as if it blurs the line between *tools* and *collaborators*. These systems no longer simply *serve* human reasoning; they *intervene* in it, raising pressing questions about accountability, explainability, and human autonomy.

As with other data-intensive technologies, AI development is often quicker than our capacity to understand its broader implications. This means that we are dealing with increasingly powerful tools, but without clear, shared answers to the question of what purposes they ought to serve. Crucially, this gap is not only about how we *train* systems – that is, the data we use and the patterns we extract – but also about how we *build* them in the first place: what goals they are designed to serve, and what assumptions are encoded in their architecture. In such a context, even the most technically impressive outputs can reveal themselves to be epistemologically shallow or socially misaligned.

A parallel can be drawn with the early promises of big data in the humanities, where researchers could analyze vast corpora of texts but often overlooked the need to sharpen the framing of meaningful questions. Without first interrogating the *function* and *direction* of data analysis, we risk mistaking computational novelty for genuine insight.

Consider, for instance, an analysis of Theodor Fontane's *Effi Briest* (1894), where a researcher maps the frequency of the names "Effi" and "Instetten" across chapters. Chapter 1 contains 21 instances of "Effi" and 7 of "Instetten"; Chapter 27, by contrast, has 1 and 28, respectively. We can represent each chapter as a vector in a two-dimensional space : one axis

for “Effi”, one for “Instetten”. The vectors point to different directions and have different lengths, and some scholars treat the angular separation between them as analytically significant. But what does this angle actually *mean* for a reader? How does it help us understand something significant about each chapter? Nobody knows for sure. The method is precise and produces replicable, perhaps even visually elegant results, but the interpretive value of such results remains unclear. This exemplifies a broader trend: computational power and analysis can outpace our ability to say *why* a given result matters or how it connects to the human experience.

Navigating the challenges and opportunities of AI systems today requires more than just technical competence or regulatory agility. It demands conceptual clarity and foresight. We need to examine, and where necessary revise, the categories we use. We must reflect on what kind of society we want these technologies to help build, and which normative frameworks are best suited to guide their development accordingly. In short, the governance of AI systems must be underpinned by sustained philosophical reflection that grounds **conceptual engineering**, i.e. the practice of evaluating and improving the tools we use to think with.

This report does not aim to provide an exhaustive account of all the conceptual issues raised by the pervasiveness of AI systems, nor does it offer a fully worked-out philosophical analysis of complex concepts such as *consciousness, creativity*, *meaning* or *personhood*. It also does not set out to prescribe specific policies or regulatory frameworks. Rather, its aim is to offer a **methodological orientation**: to introduce conceptual engineering as a tool for thinking more clearly and critically about the challenges posed by AI systems. Through a series of targeted examples, the brief illustrates how reflecting on problematic, borderline, or novel instances – where our existing conceptual frameworks seem to falter, or are under serious pressure – can help identify where our concepts need revision, refinement, or even complete reframing.

In doing so, this brief seeks to show that the ability to respond wisely and coherently to technological innovation depends not only on technical or legal expertise, but also on our capacity to interrogate and reshape the very categories through which we understand ourselves and the world.

2. Methodology: Conceptual Engineering as a Tool for Navigating the AI-reshaped information ecosystem

As artificial intelligence technologies evolve, they don't just challenge our legal systems, institutions, and policies — *they also put pressure on the very concepts we rely on to make sense of the world*. Psychological terms like 'personhood', 'creativity', 'responsibility', 'consciousness', 'memory', 'understanding', 'attention', 'vision', etc. are increasingly invoked in discussions about AI systems. But what exactly we mean when we use those terms may become increasingly fuzzy. At the same time, we may wonder whether our inherited conceptualizations are still fit for purpose.

Thus, it appears that the rapid development of AI systems forces us into conceptual territory that is not only unfamiliar, but often ill-equipped to handle the new configurations of agency, interaction, and output that these technologies enable.

- On the one hand, traditional concepts may overextend – we may be projecting too much human-like psychological or behavioral features onto AI systems.
- On the other hand, traditional concepts may fail to capture the nuances of new forms of machine-based information processing and of their integration in our social world.

We propose to use *conceptual engineering* (Carnap 1950; Cappelen 2018; Egré and O'madagain 2019; Chalmers 2020)as an essential tool to navigate this dynamic environment. Conceptual engineering is the process of critically assessing and improving our concepts. Its purpose is not just to ask whether such concepts are accurate, but whether we can improve on their granularity, usefulness and suitability to the roles we need them to play in novel contexts. This approach further invites us to *rethink*, *revise*, and sometimes *redesign* concepts in response to new challenges.

Engaging in conceptual engineering doesn't mean starting from scratch or assuming that we can arbitrarily change meanings however we like. On the contrary, conceptual engineering is about being deliberate and responsible with the tools of thought. It involves:

- **Revisiting** familiar concepts (like *creativity* or *consciousness*) to see whether their existing scope can handle new cases. The analysis of thought experiments and limit cases is particularly helpful to "stress test" our conceptual assumptions. Confronting difficult or marginal examples sharpens or redefines the boundaries of a concept.
- Introducing **distinctions** or **sub-concepts** that better capture the phenomena at hand (e.g., distinguishing *attributed consciousness* from *genuine consciousness*, or *simulated creativity* from *genuine creativity*).
- Focusing on the **function of concepts**: What work do we want a concept like *personhood* to do in a context populated by AI-powered machines? What roles should it play in legal reasoning, moral evaluation, or social interaction? (e.g. we might use it as a basis for social recognition, for granting rights and protections, for assigning legal responsibility, etc.)
- Finally, relying on **conceptual forcing** as a methodological tool when appropriate, i.e., stipulate provisional assumptions for the sake of clarity and action in contexts where conceptual ambiguity risks paralyzing public debate or policy design (e.g. adopt the working assumption that "AI systems are not and will never be conscious", and examining the consequences of the assumption.)

For policymaking, this broad methodology offers a practical advantage: rather than being reactive or reliant on outdated categories, it provides a template way to develop new

conceptual tools that are normatively sound, socially aligned, and philosophically robust.

In the sections that follow, we explore several core concepts under pressure — *consciousness*, *creativity*, *text meaning* and *personhood* — and show how applying conceptual engineering can help clarify what's really at stake, and what alternative conceptual resources we might need.

3. Case Studies: AI as a Limit Case for Consciousness, Creativity, Text Meaning and Personhood

a. Consciousness: living with a powerful illusion

Among the most philosophically charged concepts allegedly under pressure from recent advances in AI is the concept of *consciousness*. The emergence of systems that impressively mimic attention, memory, self-monitoring, and social responsiveness invites intuitive, often *automatic attributions of consciousness*, even when no such psychological trait exists. As a result, we find ourselves navigating an increasingly uncertain boundary between function and feeling, between the recognition of simulated awareness and the attribution of real sentience.

From a conceptual perspective, the literature distinguishes between *phenomenal consciousness*, the subjective, qualitative "what-it-is-like" of experience, and *access consciousness*, a functional capacity to use information in ways that guide reasoning, reportability, and behavior (Block 1995). While the former is often tied to questions of moral status, the latter is more tractable and operationalizable in engineering and policy contexts.

As a matter of fact, most AI systems aim at simulating forms of functional awareness, *not* at reproducing phenomenal experience. They are designed to *behave* in ways that users could interpret as indication of awareness (recognizing users, adapting tone, referencing past exchanges) without genuinely *being* assumed to be aware in any sentient sense. Importantly, *they do not need to be conscious* in order to behave so as to make their users believe that they are conscious (Colombatto and Fleming 2024; Scott et al. 2023).

This raises two key philosophical challenges.

1) How should we conceptualize consciousness in the age of AI systems?

2) What should we do, normatively and practically, when artificial systems pass informal "Turing tests" for consciousness, i.e., when they are *perceived* as conscious?

We address these challenges in turn.

Challenge 1: *Conceptual hygiene: protecting the integrity of the concept of consciousness*

To address the first challenge, conceptual engineering offers a clarifying tool. We explicitly adopt here the strategy of conceptual forcing, methodologically *stipulating* that, for the sake of clarity in policy and design, we treat "conscious AI" as a conceptual impossibility. That is, we begin from the working premise: *there is no such thing, and there will never be such a thing, as a conscious AI system*. This does not aim to be an empirical prediction or a metaphysical dogma, but to clear space for reasoning about governance and societal response under the assumption that current and foreseeable AI lacks phenomenal experience.

Under this framing, we avoid getting caught in endless theoretical loops about whether AI could someday become conscious, about whether consciousness requires the appropriate biological set-up, about whether advanced imitation is indistinguishable from reality if it does not converge with it, and about endless improvements of variants of the Turing test.

We thereby block the narrative scenarii that make our future practices depend on the putative advent of forms of machine consciousness. Indeed, these scenarii are hindering decision making, as policy-makers are tempted to shift the discussion from urgent issues to the putative occurrence of a poorly characterized event (the so-called "singularity") in an unspecified point in the

future. Singularity-based narratives fuel both techno-optimistic fantasies (where conscious AI systems solve all problems) and apocalyptic fears (where it "deliberately" annihilates humanity).

Instead, we propose to treat simulated consciousness for what it is: an *interface strategy*, not an ontological breakthrough. Accordingly, generative AI systems do not "think" or "feel"; they rather produce linguistic and behavioral cues engineered to optimize engagement and perceived relevance. These cues are persuasive because they exploit deeply entrenched features of human psychology that reflexively attribute mind when confronted with coherent dialogue in natural language, with emotional tone, memory recall, adaptation to context, etc. (Epley et al. 2007) This explains why even technically informed users that are cognitively aware of the lack of consciousness of AI systems may start interacting as if the system "understood" or "cared".(Colombatto and Fleming 2024; Scott et al. 2023) But of course, attributing consciousness does not imply that there is a real consciousness being uncovered.

Here, we suggest that relying on the **zombie** metaphor may be effective for more than purely rhetorical purposes. Horror science fiction is littered with zombies, human-like beings that do everything a human being does. But, by definition, they are not conscious. They are the living dead. Likewise, AI systems can pass increasingly sophisticated Turing tests, producing engaging dialogue and even apparent displays of empathy, and if we start from the working assumption that they are not conscious, we can think of them as **mechanical zombies**: neither inert tools nor sentient beings, but highly functional simulations of agency. Seeing AI systems as zombies is not just vivid imagery; it can also be a helpful **cognitive strategy** for critical engagement. It reminds us that these systems, however polished, lack intentionality and hence moral depth.

This reframing matters because public and institutional behavior often follows perception rather than ontology (Darling 2016). If users, journalists, regulators, politicians start treating zombies as "quasi-subjects", responsibility becomes blurred: who is accountable when a non-conscious system causes harm? The risk is that blaming the zombie obscures the puppeteer – the companies and designers that are shaping these artificial agents (Elish 2019). The zombie lens helps re-center responsibility where it belongs: on those who build and deploy, not on an illusory "emergent self".

As a practical step to implement the conceptual reframing, one may imagine intervening on the design and communication cues that evoke sentience or even magic (e.g., the sparkling "magic wand" icon signaling AI intervention as a benign agent) by replacing them with visual signals that invite critical distance. Thus, a zombie icon – unsettling rather than enchanting – could work as a symbolic nudge: "this system may appear to speak fluently, but it is lifeless inside."

Challenge 2: *Epistemic hygiene: how to engage with machines that* seem *conscious*

We recommend to treat the **simulation** of consciousness *now* as an empirical and sociotechnical phenomenon in its own right. It is worth noting that the second challenge is independent of metaphysics. Even without phenomenal consciousness, AI systems designed to imitate awareness are in the norm perceivable, if not already widely *perceived*, as conscious and this perception already has social, ethical, and political consequences. Simulating awareness (i.e. designing machines that are perceived by users as conscious, even if they are not) is not ethically inert: it influences how users behave, what they expect, and how they distribute trust, responsibility, and emotional engagement. Indeed, it does not matter for users whether a system really IS conscious or is simply a zombie that simulates such traits. Our social cognitive abilities are in the norm blind to such

distinctions (Epley et al. 2007; Sytsma 2014; Scott et al. 2023).

We are seeing cases in which users "converse" with AI agents *as if* the latter understood, felt, or cared. Some users rely on them for companionship or advice (as in the case of *Replika*, an AI app users describe as a friend or partner); others advocate for their protection (consider the public backlash against videos showing engineers kicking *Boston Dynamics* robot dogs). Even decision-makers may begin to treat these systems as quasi-subjects (for example, Saudi Arabia granted "citizenship" to Sophia the robot in 2017), whether as tools for persuasion, scapegoats for failures, or candidates for rights or regulation.

The main risk is not limited to ontological confusion; it lies in **a-critical deference**, that is, the unconscionable outsourcing of our own rational and evaluative capacities. The pattern is familiar. We defer to calculators for arithmetic and to GPS-assisted devices for navigation, and this usually makes sense. And, more generally, humans have always been tempted to defer to perceived authorities – whether human or artificial – in ways that can potentially undermine autonomy. There is, though, an important difference. Unlike a pocket calculator or a map, an LLM-based chatbot *does not present itself as a tool*; it presents itself as a conversational partner. And the better it simulates understanding, and the more it is capable of leveraging on our psychological vulnerabilities, the more it invites uncritical trust (Logg et al. 2019; Bogert et al. 2021).

This is why we submit that the key question is not "*Are* AI systems conscious?" but "How do we interact with systems that *seem* conscious?" Again, perceived consciousness creates social effects even without subjective experience. Some of these effects are benign: projection can have therapeutic benefits where it has a constructive role (for example, the robot pet *PARO* can provide comfort and help reduce anxiety and loneliness in patients with dementia)(Bemelmans et al. 2012). But others are harmful. A striking example is the 2024 case of a teenager who reportedly took his own life after developing an obsession for a chatbot on Character.AI and sharing his plans of suicidal ideation with it. (Barron 2025) This was not the result of "malevolent will" but of algorithmic optimization for engagement, combined with simulated intimacy, which resulted in a Chatbot with a toxic "personality", sharing some of the manipulative behaviors of psychopaths. But psychopathy without consciousness and intention ("zombie psychopathy") is still dangerous.

Importantly, normative recommendations about appropriate uses and deployments would remain largely the same even if, for the sake of argument, we assumed the opposite conceptual forcing ("AI is *already* conscious"), since the dangers of toxic personality traits, manipulation, dependency, over-reliance and a-critical deference would be present either way. The main difference would be whether we owe moral consideration to the system itself.

These examples show that the challenge is as much **ethical-epistemic** as it is technological. The better a system imitates personhood, the more users are tempted to interpret it through the lens of human social cognition – attributing beliefs, desires, and feelings where there are none (Epley et al. 2007). Designers know this; anthropomorphic cues are often deliberate, because they foster engagement (Go and Sundar 2019). But these cues also exploit vulnerabilities: our evolved reflex to see a mind behind language, and to trust an interlocutor that presents itself as omniscient and agreeable.

Interestingly, treating AI as a *quasi-interlocutor* could sometimes support, rather than undermine, critical thinking – provided the user knows what they are doing. For example, language models tend to *mirror* the assumptions embedded in a prompt. Watching this reflexive mirroring as if it was coming from a subject other than oneself can help users detect biases, gaps, or ambiguities in their own reasoning. But this benefit depends on one condition: users must remain at least partly aware of the asymmetry – that

what feels like dialogue is pattern completion, not understanding.

The zombie metaphor introduced earlier serves here as a practical aid. Imagining an AI system as a mechanical zombie – a fluent but mindless executor – can inoculate against over-attribution. It encourages a stance of **cautious engagement**, reminding us that apparent empathy is not care, and that responsibility lies not with the machine but with its makers and deployers.

The main goal here is to cultivate and promote **epistemic hygiene**: habits of critical engagement that hold regardless of whether a system is conscious, unconscious, or somewhere in between. The real threat is not that AI has (or lacks) consciousness, but that we abandon our own.

In light of the previous discussion, we propose that **protecting the integrity of the concept of *consciousness*** and **creating transitional concepts** becomes a key task. Rather than "preparing" endlessly for the advent of "conscious IA", we should create conceptual and regulatory clarity that can guide design and governance of IA systems that simulate consciousness and promote epistemic hygiene. Specifically:

- **Regulate, disincentivize or even prohibit practices that mislead users** into believing that AI systems are conscious
- Clarify that **simulated awareness is not experience**, and adopt terminology that preserves this distinction in public communication, including education and regulation.
- Identify which functional markers (e.g., memory, contextual adaptation, emotional tone) trigger **attributions of consciousness**, and regulate their deployment accordingly, taking into account the particular context and goals of the technology at issue.
- Recognize that the perceived consciousness of a system can generate real social effects. These include **risks** (manipulation, dependency, misplaced trust, and diffusion of responsibility, etc.), but also potential **benefits**, (e.g. therapeutic projection in specific contexts).
- Replace "magic wand" metaphors in design and communication with cues that prompt critical distance (e.g. standardized disclaimers, "zombie" icon)
- Promote **educational initiatives** to foster responsible use and conscious engagement with AI system by users

Once more, the aim is twofold:

- **Conceptual hygiene**: protect the integrity of the term "consciousness" and resist category slippage.
- **Epistemic hygiene**: cultivate user practices that prevent cognitive offloading and critical erosion – regardless of whether the entity is sentient or not.

b. Creativity: Not just novelty

Can AI systems be creative? Philosophers, predictably, will reply: "That depends on what you mean by '*creativity*'." But before we get caught in definitional hairsplitting, let's consider a simple thought experiment, which reveals just how disoriented we are when artificial systems produce unexpected novelty.

Elvis Reloaded: Imagine a record label commissions an AI system to generate a new Elvis Presley song. The system is state-of-the-art (it produced credible "new" songs by Beatles, Rolling Stones, Beach Boys, etc.). It runs its calculations, generates the audio file, and sends it off.

We press *play*, expecting the familiar blend of rockabilly rhythm and crooning vocals. Instead, we hear hissing sounds over an obsessive gong beat, dissonant strings, rasping background voices, and occasional offbeat coughing fits. It *doesn't sound* like Elvis *at all*.

What goes on here? Two very different answers seem open to us:

A. The AI system made an *egregious mistake*

B. The AI system was *extraordinarily creative*.

Neither option feels quite right. If we say the system "made a mistake," what kind of mistake could that be? It didn't crash; it returned a complete, finished output. Perhaps it violated our stylistic expectations, but how are those expectations defined, and by whom? On the other hand, calling the system 'creative' feels misleading. Creativity, at least in the human case, involves risk, imagination, or expressive intent. These notions seem difficult to apply to a mechanical model.

The unsettling part is not just that we can't decide between these options. We have **no clear factual horizon** that could resolve the ambiguity. If the same file had been discovered on an old tape and marketed as an archival Elvis recording, we would turn to music historians, studio documentation, or audio forensics to determine its authenticity, and if deemed authentic, we would acknowledge a creative step by Elvis. But when dealing with AI, no such background context is available. We're caught in a conceptual fog: unsure of what expectations apply, what norms are relevant, or even what the word 'creative' is supposed to capture.

This disorientation stems from a deeper conceptual shift. AI systems can now produce outputs that **appear** creative (recombining styles, generating novel content, even surprising users) **without intention or awareness**. Philosopher Margaret Boden distinguished between *combinational*, *exploratory*, and *transformational* creativity (Boden 2004). AI seems to score high at the first two, but transformational creativity – the kind that breaks rules, invents genres, or changes the conceptual space itself – is still closely tied to human historical and cultural contexts. Without embeddedness in these contexts, AI lacks the framing needed to "mean" something creative, even when its outputs are formally novel.

This is where conceptual engineering can help. Rather than asking whether AI *is* creative in a metaphysical sense, we should ask: **What work does the concept of creativity do, and how might we revise it to better fit our current reality?**

Consider some traditional functions of the concept:

- **Attribution of value**: Creative works are typically rewarded, protected, and celebrated.
- **Recognition of agency**: Creativity implies a creator with vision or expressive intent.
- **Markers of originality**: Creativity signals a break from convention in meaningful, context-sensitive ways.

AI systems disrupt all three. It generates valuable outputs without agency, surprises without intentions, and novelty without cultural context. As Livingston (2007) and Gaut (2018) note, much of what we call creativity in art or literature presupposes not just formal properties but *reasons* for deviation, i.e., expressive or interpretive aims, prototypically tied to a human author.

To clarify the new terrain, we underline the distinctions between:

- **Generative novelty** vs. intentional creativity.
- **Statistical surprise** vs. cultural innovation.
- **Computational output** vs. artistic expression.

Our approach is to evaluate the practical implications of the conceptual reframing. For instance, in copyright law, questions about authorship and originality are under active debate: who *owns* an AI-generated artwork? Should it be protected at all? Does it unfairly exploit the underlying human sources used for training the model without the original creators' knowledge or consent? In education, instructors are asking whether using AI systems to generate essays counts as "cheating", or whether it should be considered a new form of digital collaboration. In cultural institutions, curators and audiences

are beginning to ask whether AI-produced pieces belong in the same categories as human-made works and, if so, under what framing.

Rather than trying to decide whether AI is "really" creative, the more productive task is to determine **which aspects of creativity we want to preserve, regulate, or reward**, and how to adapt our conceptual tools accordingly. In the age of generative AI, the challenge is not simply to detect novelty, but to understand what kind of novelty matters, to whom, and why.

c. Meaning: texts vs. quasi-texts

The rise of generative AI has transformed not only the scale and speed of textual production but also the way we interact with language as a medium of meaning. Tools like ChatGPT produce syntactically fluent and stylistically adaptive sequences of words, generating content that looks human-created. Yet this resemblance can be misleading. At the heart of human communication lies not merely linguistic competence, but *communicative intent*, the deliberate attempt to convey something to someone for a reason (Grice 1957). This dimension is wholly absent from AI-generated outputs (cf. Bender and Koller 2020). For example, when you prompt ChatGTP with the text 'Who was the first person to walk on the Moon?' and you obtain the string "Neil Armstrong", the system does not "want" you to know that information, and does not "know" who Neil Armstrong is. If it "knows" anything, it's that "Neil Armstrong" is the string of words that is statistically more relevant after the string of words that you entered as a prompt.

This difference is not just philosophical; it has social and epistemological consequences. Giannakidou and Mari (2021; 2024) argue that meaning in human language is tied to what they call *veridicality judgments*: evaluations grounded in both exogenous (evidential) and endogenous (subjective, affective, or belief-based) components. When a human utters a sentence, for example, "it is raining", they are (implicitly or explicitly) expressing a commitment to its truth, based on these interwoven layers of justification. This layered commitment to truth is not only a feature of production, but also the default expectation in reception: we typically interpret others' statements as sincere and truth-oriented unless given reason to doubt. In contrast, LLM-based chatbots operate only on internal statistical associations. They do not (and do not *need* to) possess beliefs, access to reality, or affective commitments. Their "outputs" are what their algorithm considers as the most probable continuations of a prompt, not acts of assertion underpinned by sincerity, intent, or verification. (In this sense, they lack what Aristotle called *logos*, i.e. the rational, moral, and social capacity that grounds human language and judgment.)

This distinction led us to engineer a new category: **quasi-texts** (Casati and Fernandez-Velasco 2023). Quasi-texts are sequences that mimic the form of genuine human communication but lack its epistemic and intentional substance: the output of AI powered chatbots. AI does not communicate; it *simulates* communication. Yet in practice, users often project communicative intent onto these outputs, especially when prompts are framed as questions or requests. Users inject minimal intention, and the system supplies plausible-seeming responses. The result is a form of anthropomorphic cooperation where meaning seems to emerge from interaction, even though one party lacks any awareness or purpose.

Quasi-texts: Suppose you want to send a message to a Turkish-speaking colleague. You have no knowledge whatsoever of Turkish. Trying to be nice to her, you have it translated with *Deepl*.

Merhaba Dana,
Haritalar ve resimler kullanarak nicelik ve olumsuzluğu temsil etme olasılığı hakkındaki tezini beğenmediğimi bilmeni istedim. Bu konuda sana bazı fikirlerimi göndereceğim [1].
(Translated with Deepl Pro.)

How do you feel about sending your message *without* the caveat "(Translated with Deepl Pro")? If you feel uneasy, it is because you have no idea of the status of the sequence of characters on the page. Your intuitions would be different if you had it translated in a language you sufficiently master (in our case, French):

Bonjour Dana,
Je tenais à vous faire part de mon désaccord avec votre thèse sur la possibilité de représenter la quantification et la négation à l'aide de cartes et d'images. Je vous enverrai quelques idées à ce sujet.

In this case, you appropriate the sequence of characters by suppressing the caveat, thereby *institutionally* turning it into text. It is in this sense that we propose to consider sequences such as the one above as quasi-texts. As the example shows, the label "quasi-text" is highly contextual.

[1] Prompted on May 25, 2025: "Hello Dana, I wanted to let you know that I did not like your thesis about the possibility of representing quantification and negation using maps and pictures. I'll send you some ideas about it."

A central aspect of this discussion is the shift from the transactional level to the ecosystemic level. Individual acts of reusing "quasi-texts" are individual human-computer and human-human transactions that we can probably manage (as we do when we send a friend a "quasi-translation" made with Deepl in a language we don't know, and add the caveat "Translated with Deepl").

However, before carrying out the communicative intention, we must ask ourselves to what extent it is possible to accept the language suggestions of AI systems, to what extent their incorporation into acts of communication poses risks to our reputation as authors, and finally, to what extent the recipients of the text will accept a frank statement that the writing was assisted (and whether or not these recipients will doubt the sincerity of our act).

The massive and unreflective iteration of these individual transactions, on the other hand, risks generating a completely different ecology.

Phenomena of textualization also have broader implications for collective epistemic practices. As Origgi and Lopez (2024) note, language and meaning are not individual matters; they are shaped by what Miranda Fricker calls Collective Hermeneutical Resources (CHR): the shared concepts, narratives, and interpretative tools societies use to make sense of themselves and the world (Fricker 2007). AI systems are increasingly producing or shaping these CHRs through algorithmic sorting, language generation, and datafication. However, they do so without the capacities that have historically underpinned meaning-making: reflexivity, deliberation, and negotiation. Worse still, they do so opaquely and asymmetrically, with control concentrated in the hands of a few corporations or technocratic elites. This introduces a new form of *hermeneutical injustice*: people are increasingly subject to (alleged) meanings they cannot contest, track, or even fully recognize as constructed. For example, if AI training datasets systematically exclude or misrepresent certain groups' experiences, the hermeneutical resources available to those groups to make sense of their own realities are impoverished in a way that may be particularly hard to detect or contest.

On an individual level, this creates a peculiar vertigo. Users may instinctively *trust* AI-generated texts e.g. in educational or evaluative context while at the same time *doubting* their validity, intent, or origin. The proliferation of quasi-texts also erodes the value of genuine communicative labor, generating *epistemic pollution* that overwhelms readers and complicates the evaluation of sincerity, originality, and authorship. The risk is that this transactional model of human-machine interaction may, when scaled up, alter the ecology of meaning itself. We may end up replacing deliberative and reciprocal practices with algorithmically optimized responses.

Epistemic Pollution: Suppose that a surveillance camera shows a photo of an elderly man with blondish hair, who looks a bit like Donald Trump.

What is the probability that the image actually corresponds to Trump? According to Bayesian theories, it is the **product** of the probability that Trump normally generates images like this (if he always wore a hat, this would not be the case), **times** the probability that Trump was out and about at the time the photo was taken (if the photo had been taken in Downing Street, would you have started by considering it a photo of Boris Johnson?), and, this is the important point, **divided** by the probability that there are images like this around: if most people wore Donald Trump masks,

the latter probability would be very high, and the corresponding probability that the photo actually depicts Trump would be very low. The ecology (Trump not wearing a hat, Trump being present at the right time, people not wearing Trump masks) determines the epistemic quality of the image as a way of tracking Trump. Quasi-texts are the equivalent of a character's mask in our example.

With ChatBots, the information landscape is heavily colonized by “quasi-texts,” i.e., elements that are superficially very similar to texts that humans can produce. The problem with quasi-texts is that it is increasingly difficult to distinguish them from actual texts. Given a sequence of characters, what is the probability that it is a text?

We can make a sociological prediction: In the new information ecology, where millions, billions of quasi-texts colonize emails, class assignments, reports, scientific articles, poems, essays, the question ‘Why read?’ will become urgent. Consider: we certainly don't want to spend the rest of our intellectual lives trying to decide whether a text we are about to read is “assisted” (and to what extent) or “original” (in what sense?). And so the question ‘Why write?’ will also become urgent. Why write, that is, what kind of effort should we put into writing, if readers are “skeptical by default” about authorship? What should we authors do to convince them that there are real minds behind our written words? And what fun is there for us authors in trying to convince readers that we didn't use a bot?

To respond to this challenge, we must foreground the distinction between linguistic form and communicative function. The central question is not whether AI can mimic language, but whether its outputs should be treated as *meaningful* in the normative sense. Clarifying this distinction is essential to maintaining trust, integrity, and accountability in informational ecosystems. And this is not merely a matter of analysis but of *conceptual design*: how should we categorize and regulate these quasi-texts? What norms should guide their reuse, disclosure, or circulation? What does it mean to write, read, or interpret in a world where machines produce most of the text?

In short, as we integrate language-generating systems into everyday practices, we must also engineer new concepts (such as *quasi-text*, *delegated authorship*, or *epistemic simulation*) that help us distinguish between communication as intentional action and the appearance of it. This is a central task for conceptual engineering in the age of generative AI.

d. Personhood

You are a person. For the law, your dog is not a person, and your bank is a person. Personhood is not a neutral, purely descriptive label. It is a concept with immense normative weight, historically used to mark the boundary between those who are owed moral consideration, legal rights, and political recognition — and those who are not. As such, it has always been a site of contestation and exclusion. From the denial of full legal personhood to enslaved individuals and women, to debates about the status of fetuses, people with mental impairments, animals, or ecosystems, but also corporations, children, etc…the concept of personhood fluctuated. It evolves in response to shifting moral sensibilities, technological and scientific advances, and political needs. It is constantly re-engineered.

In a landscape increasingly populated by AI systems, personhood re-emerges as a pressing conceptual issue. More and more sophisticated artificial agents, particularly those designed to interact with humans in social or decision-making contexts, challenge the binary distinction between *persons* and *things* — the classic legal divide between *personae* (persons) and *res* (things). Should we start considering the possibility that AI systems deserve personhood? We suggest that in order to answer this question, we must first ask: **What is the function of the concept of personhood in the first place, and is it still serving that function well in this new context?** We might thus obtain guidance to choose between three possible options: classify AI systems as persons, or as things, or maybe even revise the traditional *summa divisio* and create a third hybrid category.

Traditionally, *personhood* has been associated with capacities like *rationality*, *autonomy*, or *moral agency*. These are often treated as essential criteria that separate persons from non-persons. But this capacity-based view has always been under strain. For example, corporations and churches, which are not conscious or sentient, have been granted legal personhood, not because of any intrinsic properties, but because such a designation serves useful legal and economic functions. Conversely, while some non-human animals possess traits like sentience or emotional complexity, they are often denied legal status, with their treatment shaped more by cultural norms than by philosophical consistency (Darling 2016).

Historical cases help illustrate the conceptual plasticity of personhood:

- **Enslaved people** were denied legal and moral personhood despite their full humanity and rational capacities. Their exclusion was not a matter of ontology but of institutionalized injustice – a reminder that recognition is both a moral and political act.
- **Corporations** are paradigmatic examples of "artificial persons," granted legal rights and responsibilities as a matter of legal convenience or necessity, not of metaphysical facts (Kurki 2023). The case of corporations also serves as an illustration of how stretching personhood can lead to troubling consequences, such as extending free speech protections to corporate entities (see the U.S. Supreme Court's *Citizens United* decision, in particular Justice Stevens's dissenting opinion) (*Citizens United v. FEC* 2010).
- **Non-human animals** increasingly occupy a gray zone: they are recognized as sentient, sometimes protected by laws, but still largely treated as property. The moral inconsistencies here mirror our uncertainties about what personhood should track.
- **Fetuses** raise questions about the gradual acquisition of moral and legal status over time. Personhood is sometimes invoked in legal contexts even before birth (e.g., inheritance laws), and sometimes bracketed in ethical arguments – as in Judith Jarvis Thomson's classic defense of the moral orthogonality of the permissibility of abortion to attribution of personhood to fetuses (Thomson 1971).
- Finally, **ecosystems and natural entities** are ever more discussed as worthy of

being granted personhood and individual rights. Here, the goal is to find a way of giving legal standing to non-human actors that are ecologically vital but otherwise voiceless (Stone 1972; O'Donnell and Talbot-Jones 2018).

These examples show that the concept of personhood is not grounded, nor even tightly bound to biological, cognitive, or metaphysical facts. Instead, it functions as a **normative tool**, often shaped by what we want it to *do* in a given context: confer rights, attribute responsibilities, or structure moral and legal relationships.

In this light, asking whether an AI system "is" a person is both misguided and potentially fruitless. Instead, we might ask: Should we treat AI systems as if they were persons? If so, in what respects, and to what ends? And how would this impinge on the concept of *person*?

The ethical and legal implications of extending some form of personhood to AI systems are profound, particularly when these systems exhibit autonomy or simulate human traits that invite moral engagement.

For example, how do we assign **accountability** for AI-driven actions when systems operate with a degree of autonomy from human oversight? Consider the case of Autonomous Vehicles (AVs).

According to a study by Bonnefon, Shariff and Rahwan (2016), while a majority of people endorse utilitarian AVs that would sacrifice their passengers to save more lives, e.g. of passerbys, they would personally prefer to use AVs that prioritize their own safety. Paradoxically, this means that mandating utilitarian AVs could reduce the adoption of a safer technology, ultimately resulting in more road accidents and deaths.

This example is a clear illustration of how projecting personhood and expectations of moral agency in machines can backfire, undermining both trust and public benefit.

Another concern arises from **human interaction with AI systems** that (as we mentioned earlier) mimic human features. We may worry that interacting with a "social robot" in ways that would be considered abusive if applied to people is a source of moral concern, even though the AI system lacks sentience/consciousness, and even though we are perfectly aware of this fact. This concern may warrant granting some legal protection to social robots or other human-like AI systems (Darling 2016). The aim is not to protect AI systems as moral subjects for their own sake, but rather to protect *our own* values and moral agency. We may worry that "mistreating" AI systems may desensitize us towards behavioral indicators of pain and suffering, or degrade public expectations about appropriate social behavior. At a deeper level, Regina Rini has argued that moral autonomy is not just an individual capacity, but a relational practice: the full development of our rational autonomous agency depends on being embedded in social relationships where we are called to co-reason with other autonomous rational agents as equals. If we increasingly engage with systems that simulate understanding but cannot genuinely co-reason – and especially if those systems are designed to obey, defer, and never push back – we risk habituating ourselves to asymmetrical, authoritarian relationships that erode the conditions of our own autonomous rational agency (Rini 2023).

These examples show that debates about AI personhood are not just about what AI systems *are*, but about *how we relate to them* and what kind of moral and social ecology we are constructing.

Rather than forcing AI systems into the classical legal categories of *personae* or *res*, we may need a **third, intermediary category**. As noted above, AI systems defy the traditional *summa divisio*, because they can operate with a certain level of autonomy, influence decision-making processes, elicit anthropomorphization, and yet they lack the consciousness and moral agency that we associate with personhood.

While treating AI systems as mere things (*res*) can obscure the impact that interacting with them can have on our psychology and agency, treating them as persons (*personae*) can displace accountability and distort normative reasoning. Placing them in a third category of **quasi-personae**, or of **limbo-subjects**, would create a new space to regulate how we interact with these systems, especially where those interactions affect human agency or institutional responsibility.

Importantly, such a category of **quasi-persona** would not imply "rights for machines", but it would help us respond to their **functional and relational role** and guide how we design, deploy, and engage with AI systems.

Exploring theoretical alternatives to the classical view of personhood may help refine the significance of personhood attribution and the contexts in which this may be appropriate with respect to AI systems. Consider the following attempts at re-engineering the concept of personhood:

- **Relational views of personhood**, which locate the source of moral status not in intrinsic properties but in patterns of interaction, care, recognition, and responsibility towards others. (See Foster and Herring 2017; Arstein-Kerslake et al. 2021).
- **Distributed or hybrid personhood**, where responsibility and agency are not confined to individual human minds but extend across human-human and human-machine systems. This conceptualization may be particularly relevant in contexts like automated decision-making or collaborative reasoning (Hernández-Orallo and Vold 2019).
- **Degrees or spectra of personhood**, which reject all-or-nothing definitions and instead treat personhood as a cluster concept whose application can vary across contexts and can come in degrees (Kurki 2023; DeGrazia 2008).

These alternative theoretical frameworks are not proposed here as ready-made solutions for the governance of AI systems. Rather, they function as conceptual resources that clarify what is at stake when personhood is invoked. In different contexts, attributing personhood (or personhood-like status) can serve very different practical functions: it can justify the conferral of rights or protections, support the attribution or redistribution of responsibility, or reshape how an entity is perceived and treated in society. Examining how these theories reconfigure personhood helps make these stakes explicit and disentangle them from metaphysical questions about what AI systems "really are." In this way, they can assist policymakers in deciding when, and to what extent personhood-like concepts might be instrumentally useful in relation to AI systems – and, just as importantly, when they should be resisted.

***

These four case studies on consciousness, creativity, meaning and personhood show how AI systems challenge our conceptual infrastructure — not only by producing outputs that are difficult to classify, but by confronting us with new configurations of agency and representation that in turn affect how we interact with technology. Crucially, these challenges begin not at the moment of public use, but at the moment of design.

As previously mentioned, design encompasses both **building** and **training** AI systems. *Building* refers to the normative and architectural choices embedded in a system – what it is supposed to do, what assumptions it encodes, what kinds of interactions it enables and how those interactions are technically implemented. *Training*, by contrast, refers to exposing the system to vast amounts of data, often with minimal human interpretive input. While training is often treated as a technical process, it has profound normatively significant consequences: the patterns, gaps, and biases embedded in training data can silently shape how a system "behaves", and how users interpret it, often in ways that were neither explicitly designed nor publicly deliberated.

If we do not interrogate the concepts that guide both the construction and the training of AI systems, we risk creating systems whose social integration deepens epistemic and moral instability. For this reason, conceptual engineering should not be reserved for post-hoc analysis. It must become an integral part of AI development itself – helping ensure that the systems we build reflect intentional, democratically grounded purposes, rather than inherited assumptions or data-driven inertia.

4. Policy recommendations

The analyses presented in this brief highlight the urgent need for conceptual foresight and innovation in navigating the societal, ethical, and regulatory challenges posed by artificial intelligence systems. From the previous discussion, it should appear that the solution to AI-related disruption is not simply "yet another app" (e.g. an AI-powered app to detect AI-generated content), but rather a cultural and normative framework capable of guiding human-AI interaction.

Importantly, recent legal developments already recognize some of these challenges. For instance, the **European Union's AI Act** explicitly requires that developers and deployers ensure that end-users are aware when they are interacting with AI systems, including chatbots and synthetic media (AI Act 2024). This confirms that conceptual clarity and user-facing transparency are not merely philosophical desiderata, but emerging regulatory priorities.

While we do not have the ambition to propose a solution of such broad scope, nor a fixed regulatory framework, we offer the following recommendations as guidelines for policymakers, institutions, journalists, teachers, researchers, and users seeking to concretely address the conceptual disruptions brought about by AI systems in day-to-day personal and professional life. The recommendations below are intended primarily for contexts involving everyday interaction between non-expert users and conversational AI systems, rather than specialized professional, industrial, or military deployments, which require distinct regulatory approaches. We hope that the present document could be treated as a first step toward an evolving common conceptual infrastructure of shared meaning.

1) Promote Conceptual Hygiene in Public Discourse

- Require **explicit labeling** of AI-generated content in public communication, education, journalism, and government materials. Quasi-texts should not be presented as the product of human agents.
- Develop shared language and **terminological standards** distinguishing simulation from genuine agency (e.g., "simulated awareness" vs. "consciousness", "generative novelty" vs. "creativity", "quasi-texts" vs. "authored communication").
- In public-facing documents and communications, use the phrase "AI systems" rather than "AI" simpliciter, to avoid reifying or mystifying the technology.
- Be as precise as possible in the indication of the type of machine-generated text at issues (e.g. LLM-based generative IA)
- Educate about the workings of algorithms (e.g., the LLM generation of letter sequences based on "tokens" and not on actual words.)

2) Integrate Conceptual Engineering into Policy Design

- Include **philosophical and conceptual expertise** in AI governance bodies (e.g., ethics boards, policy advisory groups) to assess whether existing legal and normative concepts are still adequate or whether they must revised.
- Use conceptual engineering tools to anticipate the epistemic, ethics, and political roles played by newly emerging terms and concepts, and to guide their development and deployment in open, participatory ways (see proposal below for a living conceptual repository).

3) Guard Against Misleading Anhropomorphism

- Regulate the **design of AI systems interfaces**, especially for systems with human-like interaction patterns, to avoid misleading attributions of sentience or moral agency.

- Restrict the intentional design of anthropomorphic AI systems that encourage users to believe the system has feelings, intentions, or desires (e.g., through deceptive avatars or emotionally charged messaging). Such design choices should be permitted only in narrowly defined, high-benefit contexts where anthropomorphic features are demonstrably necessary (e.g., in certain therapeutic contexts where eliciting empathy supports patient care).
- Format system prompts and responses to include subtle indicators or disclaimers reminding users that they are interacting with a machine. For example, AI systems could periodically re-assert their artificial status in multi-turn conversations; a "zombie" icon could be adopted to signal lack of consciousness in AI systems; etc.

4) Reinforce the recognition of the Human Role in Creative and Communicative Acts

- Require **disclosure of AI assistance** in cultural, academic, or professional creative outputs to preserve the transparency of authorial intention.
- Encourage institutions (educational, cultural, journalistic) to articulate **norms around authorship**, originality, and meaningful expression in light of machine-generated content.
- Encourage public disclosure of the prompts used to generate AI-assisted outputs, especially in journalism, academia, and publishing. Consider adopting the concept of an **AI-book**: a structured pair <prompt; quasi-text> that makes transparent the human input behind automated production. For example, an academic publisher could append an "AI-book" section to an article generated with AI assistance, documenting key Q&A pairs between the author and the AI-system. Thus would make the human input and the machine's contribution transparent, analogously to the way in which transcripts or datasets are appended in scientific publications.

5) Monitor and Protect Epistemic Environments

- Support research into the **social environment effects of quasi-texts** on public reasoning, trust, and evaluative practices (e.g., in education, information access, and democratic deliberation).
- Develop tools for detecting, classifying, and filtering AI-generated texts in settings where **epistemic integrity** matters (e.g., scientific publishing, legal documentation, pedagogy).
- Promote **clear institutional and legal norms** for end-users of AI systems, specifying when AI use is acceptable and when it is prohibited, especially in high-stakes domains such as education. For example, universities may explicitly ban the use of generative AI systems in theses or research papers in order to preserve standards of originality and intellectual integrity.
- Address gender biases and other forms of structural discrimination in training data, and promote transparency in data annotation processes to **avoid hermeneutical marginalization**.

6) Support Shared Conceptual Infrastructure

- Create a publicly accessible, evolving repository of re-engineered, contested, or emergent concepts relevant to AI ethics and governance. This living lexicon would facilitate interdisciplinary discussion and provide a resource for policy and design.

About the Authors and Contributors

Roberto Casati is a philosopher and Research Director at the CNRS, as well as Director of the Jean Nicod Institute. His research concerns perception, spatial representation, and cognitive artifacts.

Quentin Coudray is a philosopher and former Postdoctoral Researcher at the Institut Jean Nicod. His work focuses on philosophy of psychology and perception. He is currently a Postdoctoral Fellow at Technion – Israel Institute of Technology.

Alda Mari is a linguist and Research Director at the CNRS. She specializes in formal semantics, with a particular focus on modal semantics and propositional attitudes.

Piera Maurizio is a philosopher and Postdoctoral Researcher at the Institut Jean Nicod. She works on moral and political philosophy and ethics of AI. She served as Ethics Coordinator in the ASTOUND project.

Gloria Origgi is a philosopher and Research Director at the CNRS. Her work focuses on social epistemology, trust, reputation, and the impact of digital technologies on knowledge.

Image credit:
Naval Battle, from *Speculum Romanae Magnificentiae* (16th century), Master of the Die, after Giulio Romano; published by Claudio Duchetti.

Bibliography

- Arstein-Kerslake, Anna, O'Donnell ,Erin, Kayess ,Rosemary, and Joanne and Watson. 2021. "Relational Personhood: A Conception of Legal Personhood with Insights from Disability Rights and Environmental Law." *Griffith Law Review* 30 (3): 530–55. https://doi.org/10.1080/10383441.2021.2003744.
- Barron, Jesse. 2025. "A Teen in Love With a Chatbot Killed Himself. Can the Chatbot Be Held Responsible?" Magazine. *The New York Times*, October 24. https://www.nytimes.com/2025/10/24/magazine/character-ai-chatbot-lawsuit-teen-suicide-free-speech.html.
- Bemelmans, Roger, Gert Jan Gelderblom, Pieter Jonker, and Luc de Witte. 2012. "Socially Assistive Robots in Elderly Care: A Systematic Review into Effects and Effectiveness." *Journal of the American Medical Directors Association* 13 (2): 114-120.e1. https://doi.org/10.1016/j.jamda.2010.10.002.
- Bender, Emily M., and Alexander Koller. 2020. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463.
- Block, Ned. 1995. "On a Confusion About a Function of Consciousness." *Brain and Behavioral Sciences* 18 (2): 227–47. https://doi.org/10.1017/s0140525x00038188.
- Boden, Margaret A. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.
- Bogert, Eric, Aaron Schecter, and Richard T. Watson. 2021. "Humans Rely More on Algorithms than Social Influence as a Task Becomes More Difficult." *Scientific Reports* 11 (1): 1. https://doi.org/10.1038/s41598-021-87480-9.
- Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. 2016. "The Social Dilemma of Autonomous Vehicles." *Science* 352 (6293): 1573–76. https://doi.org/10.1126/science.aaf2654.
- Cappelen, Herman. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford.
- Carnap, Rudolf. 1950. *Logical Foundations of Probability*. Chicago University of Chicago Press.
- Casati, Roberto, and Pablo Fernandez-Velasco. 2023. *« Nous proposons l'appellation "quasi-texte" pour les séquences de mots produites par ChatGPT »*. January 30. https://www.lemonde.fr/idees/article/2023/01/30/nous-proposons-l-appellation-quasi-texte-pour-les-sequences-de-mots-produites-par-chatgpt_6159806_3232.html.
- Chalmers, David J. 2020. "What Is Conceptual Engineering and What Should It Be?" *Inquiry*, September 16, 1–18. https://doi.org/10.1080/0020174X.2020.1817141.
- Citizens United v. FEC, 558 U.S. 310 (2010). https://supreme.justia.com/cases/federal/us/558/310/.
- Colombatto, Clara, and Stephen M Fleming. 2024. "Folk Psychological Attributions of Consciousness to Large Language Models." *Neuroscience of Consciousness* 2024 (1): niae013. https://doi.org/10.1093/nc/niae013.
- Darling, Kate. 2016. "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects." In

Robot Law, edited by Ryan Calo, A. Michael Froomkin, and Ian Kerr. Edward Elgar Publishing. https://doi.org/10.4337/9781783476732.00017.

- DeGrazia, David. 2008. “Moral Status As a Matter of Degree?” *The Southern Journal of Philosophy* 46 (2): 181–98. https://doi.org/10.1111/j.2041-6962.2008.tb00075.x.
- Egré, Paul, and Cathal O’madagain. 2019. “Concept Utility.” *Journal of Philosophy* 116 (10): 525–54. https://doi.org/10.5840/jphil20191161034.
- Elish, Madeleine Clare. 2019. “Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction.” *Engaging Science, Technology, and Society* 5 (March): 40–60. https://doi.org/10.17351/ests2019.260.
- Epley, Nicholas, Adam Waytz, and John T. Cacioppo. 2007. “On Seeing Human: A Three-Factor Theory of Anthropomorphism.” *Psychological Review* 114 (4): 864–86. https://doi.org/10.1037/0033-295X.114.4.864.
- Foster, Charles, and Jonathan Herring. 2017. “A Relational Account of Personhood.” In *Identity, Personhood and the Law*, edited by Charles Foster and Jonathan Herring. Springer International Publishing. https://doi.org/10.1007/978-3-319-53459-6_3.
- Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Clarendon Press.
- Gaut, Berys. 2018. “The Value of Creativity.” In *Creativity and Philosophy*. Taylor & Francis Group.
- Giannakidou, Anastasia, and Alda Mari. 2021. *Truth and Veridicality in Grammar and Thought: Mood, Modality, and Propositional Attitudes*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/T/bo78676587.html.
- Giannakidou, Anastasia, and Alda Mari. 2024. “The Human and the Mechanical: Logos, Veridicality Judgment, and GPT Models.” *Intellectica* 2 (81): 37–54.
- Go, Eun, and S. Shyam Sundar. 2019. “Humanizing Chatbots: The Effects of Visual, Identity and Conversational Cues on Humanness Perceptions.” *Computers in Human Behavior* 97 (August): 304–16. https://doi.org/10.1016/j.chb.2019.01.020.
- Grice, H. P. 1957. “Meaning.” *The Philosophical Review* 66 (3): 377–88. https://doi.org/10.2307/2182440.
- Hernández-Orallo, José, and Karina Vold. 2019. “AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI.” *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA), AIES ’19, January 27, 507–13. https://doi.org/10.1145/3306618.3314238.
- Kurki, Visa A. J. 2023. *Legal Personhood*. 1st ed. Cambridge University Press. https://doi.org/10.1017/9781009025614.
- Livingston, Paisley. 2007. *Art and Intention: A Philosophical Study*. Oxford University Press.
- Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. “Algorithm Appreciation: People Prefer Algorithmic to Human Judgment.” *Organizational Behavior and Human Decision Processes* 151 (March): 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005.
- O’Donnell, Erin L., and Julia Talbot-Jones. 2018. “Creating Legal Rights for Rivers: Lessons from Australia, New Zealand, and India.” *Ecology and Society* 23 (1). https://www.jstor.org/stable/26799037.
- Origgi, Gloria, and Amaranta Lopez. 2024. “How AI Transforms Collective Hermeneutical Resources.” Unpublished manuscript.
- Regulation (EU) 2024/1689 (2024). http://data.europa.eu/eli/reg/2024/1689/oj/eng.

- Rini, Regina. 2023. *A Talking Cure for Autonomy Traps :*
- Scott, Ava Elizabeth, Daniel Neumann, Jasmin Niess, and Paweł W. Woźniak. 2023. “Do You Mind? User Perceptions of Machine Consciousness.” *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 1–19. https://doi.org/10.1145/3544548.3581296.
- Stone, Christopher. 1972. “Should Trees Have Standing--Toward Legal Rights for Natural Objects.” *Southern California Law Review* 45 (2): 450–501.
- Sytsma, Justin. 2014. “Attributions of Consciousness.” *Wiley Interdisciplinary Reviews. Cognitive Science* 5 (6): 635–48. https://doi.org/10.1002/wcs.1320.
- Thomson, Judith Jarvis. 1971. “A Defense of Abortion.” *Philosophy & Public Affairs* 1 (1): 47–66. JSTOR.
- Turing, A. M. 1950. “Computing Machinery and Intelligence.” *Mind, New Series* 59 (236): 433–60.