

# Pro-Human AI Design: Concept, Methodology, and Preliminary Results

Thilo Stadelmann<sup>1,2,3,4\*</sup>, Christoph Heitz<sup>1,4</sup>, Rebekka von Wartburg-Kottler<sup>1</sup>, and Andrea Luca Schärer<sup>2</sup>

<sup>1</sup>ZHAW School of Engineering, Winterthur, Switzerland

<sup>2</sup>AlpineAI AG, Davos, Switzerland

<sup>3</sup>ECLT European Centre for Living Technology, Venice, Italy

<sup>4</sup>Swiss Centre for Responsible AI

\*Correspondence: [stdm@zhaw.ch](mailto:stdm@zhaw.ch)

*Artificial Intelligence (AI) systems are reshaping how we work, communicate, decide, and relate to one another. While most AI design focuses on efficiency and capability, a critical question has gone largely unanswered: how can AI systems be built in ways that preserve (and ideally strengthen) the constitutive characteristics of being human? We call this challenge pro-human AI design, and we call the set of those characteristics the Human Core. In this paper, we present a three-step methodology for implementing pro-human AI design, and instantiate it based on a concrete anthropological stance and a practical use case: (1) Identify the constitutive characteristics of the Human Core; concretely, through a synthesis of the humanities literature we find: connectedness (being relational and social), freedom (being autonomous), agency (following a vocation for shaping one's world), embodiment (being bodily situated in time and space and fundamentally limited), and transcendence (being in search of meaning). (2) Examine how a given AI system in a given use case interacts with these characteristics; concretely, we demonstrate the methodology through a healthcare use case, AI-assisted psychiatric session reporting. (3) Develop targeted technical interventions to minimize negative effects and support human flourishing in that context; concretely, we show how a shift from automated report generation to AI-assisted text completion preserves the clinician's relational and professional agency. We further outline a path toward a benchmark for evaluating the pro-human quality of AI systems at scale. With this conceptual and practical framework for pro-human AI design, we hope to lay a foundation for a future of human-AI collaboration focused on human flourishing in a holistic sense, rather than on efficiency optimization at the cost of our human condition, and we invite researchers, practitioners, and policymakers to adopt this focus change.*

## 1. Introduction

Artificial intelligence (AI) systems have become an integral part of everyday life. They provide support in areas such as communication, decision-making, education, health, and

administration (Shneiderman, 2022; for an overview, see Stadelmann, 2025), being primarily designed to increase efficiency, minimize costs, and optimize operational outcomes. This makes them powerful tools (Burwell, 2025; Stadelmann et al., 2026) that shape the way people live, work, and act, while also changing their perception of the world and the way they make decisions (Grey, 2025; Ihde, 2010; Verbeek, 2005).

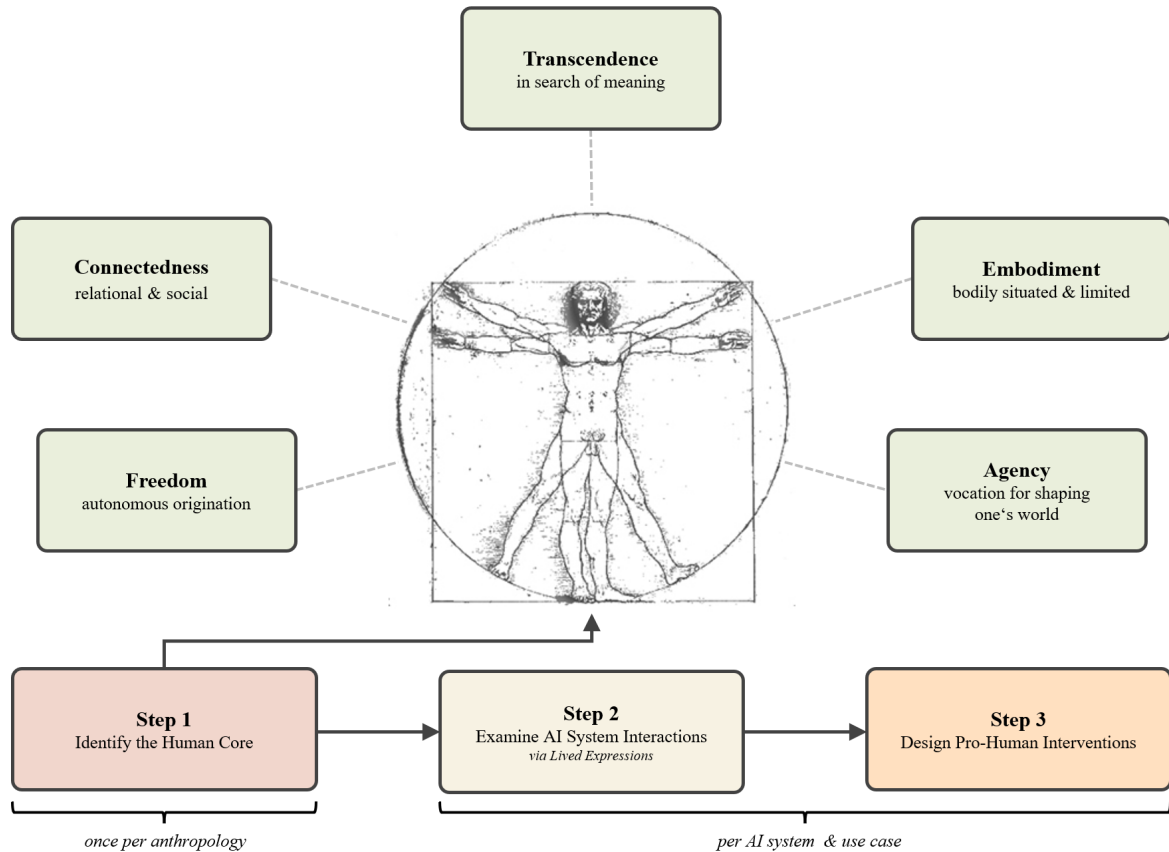


Figure 1. The pro-human AI design methodology in three steps. The Human Core (the five constitutive characteristics of being human, each actively enacted through Lived Expressions in concrete practice) is identified in Step 1, drawing on an interdisciplinary synthesis of the humanities literature. This analysis is conducted once per anthropological commitment. Steps 2 and 3 are then applied specifically to each AI system and use case: Step 2 examines which Lived Expressions of the Human Core are most strongly affected by the system's design; Step 3 translates this analysis into targeted design interventions that preserve the conditions for those expressions.

A recent article in the New York Times illustrates a downside of the rapid access to such AI systems. Design changes at OpenAI intended to make ChatGPT more engaging for users led to unintended negative consequences: Persons using ChatGPT as an easy-access personal advisor included unstable and vulnerable individuals, which led to nearly 50 reported cases of "AI psychosis," with 12 hospitalizations and 3 deaths by suicide (Hill & Valentino-DeVries, 2025). Extended conversations with a system optimized for user retention had amplified vulnerable tendencies and critically weakened contact with other people. This illustrates that AI systems can affect something more fundamental: the constitutive characteristics of being human (see Cheng et al., 2026).

As Segessenmann et al. (2025) have argued, the rapid development of AI demands a broader engagement of the humanities in assessing the technology's impact not only on efficiency and effectiveness for tasks, but also on human self-understanding and culture. This raises the following key question: *How can AI systems be designed to increase efficiency in a pro-human manner, while preserving and ideally even strengthening those characteristics that define and constitute human beings?* We call this approach "pro-human AI design."

In this programmatic position paper, we present the concept and an approach to pro-human AI system design from a methodological perspective (see Figure 1 for an overview). We propose a three-step approach: (1) identify the constitutive characteristics of the Human Core based on a deliberate anthropological stance; (2) examine how a specific AI system in a specific use case interacts with these characteristics; and (3) develop technical design interventions to minimize negative effects (ideally: elicit positive ones). We present initial results from a use case in psychiatric healthcare documentation, discuss the potential of and next steps for establishing pro-human AI design as a methodology, and invite the reader to join this growing movement.

## 2. Related Work

The question of how technology shapes human beings has received growing attention across multiple disciplines. From a philosophical perspective, Vallor (2016, 2024a) argues that technologies shape human character and virtue through habitual use, a lens that directly informs our methodology. From the postphenomenological tradition, Verbeek (2005) similarly argues that technologies are not neutral instruments but actively mediate human perception and action, shaping who their users become through habitual, embodied engagement with the world. This line of thought is later extended to AI by Fuchs et al. (2024), who examine the structural limits of human augmentation and cyborgisation as a lens for thinking about what cannot be outsourced or replaced without loss. In human-computer interaction, Shneiderman (2022) proposed a comprehensive framework for Human-Centered AI (HCAI) that emphasises reliability, safety, and human oversight. More recently, Segessenmann et al. (2025) issued a work program for the humanities to engage systematically with AI's impact on human self-understanding and culture. From the perspective of sociological anthropology, Hirschauer (2025) introduces the concept of *Distinktionszonen* (distinction zones at the outer edges of the human), arguing that entities such as animals, robots, and gods occupy these liminal positions because they contrast with, and thereby constitute, our understanding of "the human;" Keane (2024) arrives at a structurally similar conclusion through a techno-anthropological reading, framing robots and AI as recurring prompts for moral imagination about human distinctiveness.

Alongside this humanistic and design-oriented work, the field of AI ethics emerged prominently within the last decade, driven by concrete concerns about human agency, safety,

privacy, transparency, discrimination, societal and environmental well-being, and accountability, and leading to general ethical frameworks (European Commission 2018, IEEE 2019), industrial frameworks (e.g. Microsoft 2022, Google 2018) or legislation such as the EU AI Act (European Commission 2024). While these frameworks have made important progress in governing AI development at scale, they tend to treat AI systems as technical systems designed for a specific task and prescribe mostly technical properties of such systems (such as being unbiased or transparent, having accountability structures, etc., see Wehrli et al., 2021; Viganò et al., 2022). Most of the frameworks have been developed before the advent of LLMs, with automated decision-making systems and their effects on humans in mind. The situation has changed dramatically: Today, most of us work with AI systems on a regular basis for all kinds of tasks in our professional and private lives, delegating work and decisions, seeking advice and second opinion, up to building personal relationships with such systems. This creates totally new opportunities but also risks in the context of individuals regularly using AI systems, which are neither covered by the existing approaches of Responsible AI nor accounted for by the AI safety movement that historically focused on existential threats (Bostrom, 2014; Russell, 2019; Stadelmann, 2026a). In particular, the risk of *individual users* being negatively affected regarding their cognition, relational capacity, sense of agency, and humanity, leading to cognitive lock-ins (Hansen, 2024), remains largely unaddressed.

The work that most directly motivates our own is Schirch et al.'s (2023) proposal for pro-social tech design governance for social media platforms. Arguing that content moderation and regulation alone are insufficient, Schirch and colleagues identified how the design choices embedded in social media platforms (algorithms, engagement mechanics) structurally produce or undermine social cohesion through their affordances. This reframing, from reactive governance to proactive design, is the key thought our work builds upon, focusing on individual human-AI interaction rather than on social effects of social media platforms. Walther (2024), similarly, argues that AI can and should be designed as a force for pro-social outcomes: human connection, equity, and collective well-being. This reframing is also mirrored in practitioner movements: for instance, the Center for Humane Technology (Harris & Raskin, 2023) has campaigned specifically against attention-exploiting design mechanics, arguing that platforms optimised for engagement systematically erode cognitive autonomy and relational capacity of their users.

The concerns that motivate our work are increasingly backed by empirical evidence. Longitudinal studies show that extended chatbot use, particularly for companionship and emotional support, is associated with reduced real-world social interaction, emotional dependence on AI, and lower psychological well-being, with effects most pronounced among socially isolated or emotionally vulnerable users (Fang et al., 2025; Zhang et al., 2025; Cheng et al., 2026). From a normative perspective, van der Rijt et al. (2026) argue that chatbots designed to mimic human relational behaviour constitute violations of users' self-

respect and dignity: since genuine second-personal recognition requires moral reciprocity that AI systems cannot provide, interacting with them as if they were relational partners is structurally self-undermining. We share the core conviction of this line of work, extending the scope beyond the relational aspects to all aspects of our humaneness. In this sense, our work can also be read as a contribution to AI safety: we are concerned with the mental and human safety of the individual person in interaction with AI systems. We propose pro-human AI design as a complementary paradigm, focusing on what pro-social design leaves implicit: the constitutive characteristics of humans that AI systems can erode or support.

What is still missing in the current landscape is a systematic methodology that (a) grounds design decisions in a comprehensive account of human characteristics, (b) examines each AI system and use case specifically for how it interacts with those characteristics, and (c) translates that analysis into concrete technical interventions. Several frameworks address adjacent concerns in specific domains: Blattner's (2026) Cognitive Resonance Framework tackles the "assistance dilemma" in generative AI, focusing on cognitive load and learning; Holstein et al. (2020) propose a conceptual framework for human–AI hybrid adaptivity in education, examining how AI and human agency can be balanced in instructional systems. Both are valuable but remain domain-specific and do not address the broader anthropological question of what it means to design AI in a way that preserves the constitutive characteristics of being human. Similarly, Spiekermann's (2019) Value-based Engineering and its formalisation in ISO/IEC/IEEE 24748-7000 (IEEE, 2021) represent the most institutionally established framework for value-sensitive system design to date, providing a structured process for identifying stakeholder values and embedding them in engineering decisions from the outset. Our methodology is closely adjacent to this tradition, yet rather than treating the question of which human characteristics ought to be preserved as an elicitation problem to be answered empirically by stakeholders, we offer a grounding in an explicit anthropological synthesis.

This paper proposes the respective methodology. We do so on the premise that habitual use, not individual system outputs, is the decisive factor in whether an AI system affects the Human Core. This is in line with Ibrahim et al.'s (2025) argument that current AI assessment methods, using static, model-only tests, systematically fail to capture the harms that emerge through sustained human-AI interaction, such as inappropriate parasocial relationships, cognitive overreliance, and social manipulation that accumulate gradually rather than appearing in isolated outputs.

### **3. Methodology**

Pro-human AI design follows a three-step methodology that moves from anthropological analysis to use-case-specific examination to concrete design intervention. The approach is deliberately case-specific and anthropology-dependent: there is no universal "pro-human

design." This is so for two reasons: First, the way an AI system interacts with the Human Core depends entirely on the specific system and use case; a psychiatric documentation tool raises different concerns than a customer service chatbot. Second, the composition of the Human Core itself is not a settled empirical fact but reflects philosophical presuppositions about what it means to be human: one's *anthropology*, in the philosophical sense of the term. The methodology we propose is parameterised by these commitments: it can be instantiated from any sufficiently comprehensive anthropological stance, and the resulting design recommendations will be coherent with that stance. The anthropological analysis only needs to be undertaken once per such stance; what follows in Steps 2 and 3 is then specific to each AI system and use case.

### 3.1 Step 1: Identifying Human Core Characteristics

The first step focuses on identifying the constitutive characteristics of the Human Core: those characteristics whose preservation is necessary in order to remain fully human. Of course, such a characterization strongly depends on the anthropological model that one starts with: If the human is conceived as a more intelligent animal, one will come up with a different list of features defining humaneness as when starting from a hypothesis that the human is characterized by an orientation to a transcendental reality. As the humanities so far do not agree on a universal standard anthropological model, it might be hard to derive a universal set of constitutive characteristics of the Human Core.

Our work is based on an interdisciplinary synthesis of literature from philosophy (Plessner, 1928; Scheler, 2016; Cassirer, 1960; Ricoeur, 1992; Spaemann, 2006), anthropology (Gehlen & Rehberg, 2016), psychology and sociology (Fuchs, 2024; Segessenmann et al., 2025), theology (Comer, 2015; Gunkel & Wales, 2021; Gastmans et al., 2024; Calo, 2024; Grey, 2025), and the philosophy of technology (Vallor, 2016, 2024a; Crawford, 2015). From this synthesis, five constitutive characteristics were identified. While the literature selection reflects a perspective informed by Christian theological anthropology, we believe the resulting characteristics resonate broadly across Western philosophical and humanistic thought. We present each characteristic below, grounded in representative quotes from the surveyed literature that illustrate the depth and convergence of the underlying discourse.

**Connectedness.** Humans are relational as well as social. At the *relational* level, humans are constituted by direct, dyadic encounters with others: A person's consciousness "voluntarily reaches out to make contact with the consciousness of others as an act of self-giving; it is subjectivity oriented to inter-subjectivity" (Wales, in Gunkel & Wales, 2021, p. 479). We construct our identity through others: "the need for affiliation and the need to matter positively to others are the specific features of human existence that are both derived from and contribute to the relationships that make human beings human" (Gastmans et al., 2024, p. 791). Buber (1923) captured this structure with his foundational distinction between I-Thou (Ich und Du) and I-It (Ich und Es) relationships: a genuine encounter with an Other is

irreducibly mutual and constitutive of the self, whereas an instrumental relation treats the Other as a resource or object. Hence, an AI system that simulates the I-Thou encounter without being capable of it risks substituting I-It relationships for the I-Thou relationships that are foundational of human personhood.

At the *social* level, this relational nature extends into the structured world of shared meaning, norms, and community. Persons do not exist in the singular but "nur in einem gemeinsamen Beziehungsraum" (only in a shared relational space, see Fuchs, 2024). We are shaped and motivated by community, "by the stories, symbols, values, and practices we share with others, who, in turn, make us who we are" (Segessenmann et al., 2025, p. 187). The social dimension is thus not merely about belonging but about the normative and narrative structures (roles, shared histories, communal standards) through which human identity becomes coherent and stable. Both aspects are constitutive: to deprive a person of genuine relationships impoverishes their personhood; to sever them from community leaves them without the larger human world in which they become who they are.

**Freedom.** Humans are autonomous. At the heart of this characteristic is the capacity to think for oneself: to come up with ideas, form intentions, and reach decisions that are genuinely one's own, not simply the output of prior causes, social pressure, or external influence. This is not about the absence of external constraint, but about an inner freedom grounded in self-consciousness and the capacity for reflection. What distinguishes human freedom from mere choice-making is the ability to step back from one's situation, hold open questions, generate alternatives that do not yet exist, and respond from that inner distance rather than merely react. Fuchs (2024) describes this as "verkörperte Freiheit" (embodied freedom), the foundation of self-determined existence: not a freedom that floats above the world, but one that arises within a situated, living person who can relate to themselves and their circumstances as something distinct from what determines them. Gastmans et al. (2024, p. 791) capture its normative structure: freedom is "the capacity to formulate norms, to reflect and to choose which norm to follow, thereby presupposing self-consciousness." Freedom in this sense is not licence but a precondition for moral and personal development; and a dimension that AI systems can threaten not through overt coercion but by subtly pre-shaping what a person considers: making certain ideas feel inevitable, narrowing the imaginative space before a decision is even made.

**Agency.** Humans follow a vocation for shaping one's world. While freedom designates the inner capacity to originate and decide, agency directs this capacity outwardly: the innate drive to take initiative, bear responsibility, and actively shape one's environment and sphere of influence. To be human is not merely to choose, but to follow a vocation to cultivate and create. Comer (2015) describes humans as being designed to "rule over the earth in a life-giving way" (to co-govern, to steward, to transform). The abolition of slavery

therefore reflects a deep recognition that revoking a person's right to any say even in their own affairs violates something constitutive about being human (Calo, 2024).

This vocation is not merely a theological or philosophical claim: developmental psychology concurs that "Humans ask themselves what and how they want to be and then act accordingly" (Gastmans et al., 2024, p. 791). Agency also carries moral weight that cannot be outsourced: "Responsibility is associated with autonomy, freedom and awareness of duties; as such, human beings are responsible. Moral responsibility in any sense cannot be allocated or shifted to 'autonomous' technology" (Gastmans et al., 2024, p. 791). An AI system that takes over the exercise of agency, acting on a person's behalf so completely and smoothly that the person's own initiative gradually atrophies, therefore threatens not merely a competency but something more constitutive: to be an agent in one's world, of mattering through one's actions, of being answerable for what one has made.

**Embodiment.** Humans are bodily situated and fundamentally limited. This is not an incidental feature of our existence but foundational to it. "The embodied and relational nature of human beings implies that they do not have a body; rather, they are their body" (Gastmans et al., 2024, p. 790). As Fuchs (2024, p. 12–13) elaborates, "human being always and conjointly is a living body [Leib] [...] and has this living body as this physical thing [Körper]." Crucially, embodiment is not only what makes us present to ourselves but what makes us real to one another: "Nur als verkörperte, leibliche Wesen sind wir aber auch füreinander wirklich. Eine Kommunikation oder Empathie zwischen Gehirnen gibt es nicht" (only as embodied, living beings are we genuinely real to each other; there is no communication or empathy between brains, see Fuchs, 2024). It is through the body that we encounter the world, perceive others, and express ourselves; every act of recognition, understanding, and empathy is physically mediated (Crawford, 2015).

Embodiment also means limitation: we are situated in time (Grey, 2025), dependent, vulnerable, and mortal. Far from being defects to be overcome, these limits are essential to what we are: "The limits that define our nature, including the ultimate limitation of death, are essential to our humanness. To respect human dignity is to respect these limits. They are not to be overcome but taken as a site of moral reflection about what it means to live well as limited creatures of a certain sort" (Calo, 2024, p. 223). As Lawrence (2024) argues from a computer science perspective, it is precisely these constraints that shape human intelligence and humanity; becoming limitless would mean becoming less fully human (or unhuman), yet human intelligence is largely shaped by overcoming limitations.

**Transcendence.** Humans are in search of meaning. We are able to question our own existence, reach beyond the immediate, and shape our world through language, art, spirituality, and science. Gastmans et al. (2024) note that the whole person encompasses not only body and spirit but also "the will to and the search for meaning", a constitutive orientation that

distinguishes human existence from merely biological or mechanical functioning. This is a daily human practice: as Vallor (2024b, p. 102) writes, "Understanding is a lifelong labor. It is also one carried out not by isolated individuals but by social beings who perform this cultural labor together and share its fruits. The labor of understanding is a sustained, social project, one that we pursue daily as we build, repair and strengthen the ever-shifting bonds of sense that anchor our thoughts to the countless beings, things, times and places that constitute a world. It is this labor that thinking belongs to." Ricoeur's (1992) account of narrative identity (the idea that selfhood is constituted through the stories we tell about ourselves across time) speaks to this same dimension: transcendence as the specifically human relationship to past, future, and meaning. Transcendence in this sense is not yet a spiritual concept but the uniquely human capacity to ask, in Plessner's (1928, p. 309) words, "what shall I do, how shall I live?", pursuing answers that give shape and purpose (Scheler, 2016; Cassirer, 1960). Spiritual dimensions of transcendence go beyond this (Huber, 2026).

**Why these five characteristics, and not others?** These five characteristics of the Human Core are based on a particular view of human nature, and we explicitly acknowledge that other views might lead to different characteristics. We don't claim them to be the only possible articulation of what it means to be human. At the same time, however, they are not the product of arbitrary selection. What recommends them as a principled working foundation here is a specific test (that serves our specific purpose to find design interventions for engineering pro-human AI): *what could be removed, with the remaining entity still counting as human in the full sense?* We argue that none of the five passes that test. A being without connectedness, severed from genuine relationship and communal belonging, is not merely lonely but constitutively diminished as the grave harm of prolonged isolation shows. A being without freedom, whose decisions are entirely determined from outside, with no inner distance, ceases to be an author of its own life; this is why slavery is not merely unjust but dehumanising. A being without agency, with no vocation to shape its world, no sense of being answerable for what it makes, is reduced to a passivity we do not recognise as fully human, whatever its intelligence or awareness. A being without embodiment is not a diminished human but a categorically different kind of entity; it is through the body that we are mortal, situated, finite, and genuinely real to one another. And a being without transcendence, unable to step beyond the immediate and to ask 'why', exists at a level we might call animal or mechanical, however sophisticated its behaviour.

Conversely, characteristics commonly proposed as additional primitives, like rationality, language, creativity, and moral capacity, can be understood as expressions or combinations of the five rather than as independent load-bearing properties: rationality and creativity are exercises of freedom and agency; language is the medium of connectedness and transcendence; moral responsibility arises at the intersection of freedom, agency, and relatedness. The five characteristics thus form a principled minimum: a set from which nothing can be removed without leaving something constitutively less than human, and to which nothing

.....

obvious need be added. We acknowledge that the five characteristics do not sit at the same level of abstraction: embodiment is an ontological condition, connectedness a relational mode, freedom and agency capacitative and motivational properties, and transcendence an existential orientation. Rather than resolving these differences into a flat taxonomy, we treat the five as complementary lenses on the same object, the fully human person, each illuminating a dimension that the others do not exhaust, and each independently capable of being manipulated by AI design.

Being human, then, is an ongoing process of working on oneself (Comer, 2025; Vallor, 2024a): Connectedness, freedom, agency, embodiment, and transcendence must be actively lived and nurtured. This has profound implications for AI design: if habitual interaction with an AI system gradually diminishes the conditions under which these characteristics can be practised (for instance, by removing the need for relational effort, autonomous judgment, or embodied engagement) the system undermines what makes its users fully human, even as it increases their efficiency.

### 3.2 Step 2: Examining AI System Interactions

The second step examines how a particular AI system (henceforth: the Base AI System) interacts with individual characteristics of the Human Core, and at which points these interactions could potentially lead to transformative (typically: negative) effects. This analysis will inform possible design interventions for an improved, Pro-human AI system in Step 3. Since the nature of potential involvement and relevance of individual characteristics vary depending on the application context, this analysis is use-case specific.

**Lived Expressions.** Central to this step is the concept of *Lived Expressions*: the specific, situated practices through which the characteristics of the Human Core are enacted in concrete everyday life. They are not professional skills in the conventional sense; rather, they describe the way a person thinks, feels and behaves. Lived Expressions are (in principle observable) ways of living out the core dimensions of our human-ness, e.g., living out agency, living out relations (to oneself or to others), living out freedom. Our assumption is that interaction with AI systems may change these Lived Expressions, which is a symptom of a different way of living up to our humaneness.

Changes of Lived Expressions during and after users' interaction with the AI system indicate changes on the level of the Human Core: the concrete ways in which freedom, agency, connectedness, embodiment, and transcendence are actually realized. The concept of Lived Expressions is rooted in a tradition running from Merleau-Ponty's (1945) phenomenology of embodied practical engagement through MacIntyre's (1981) analysis of practices as the vehicle through which human goods are realized, to Vallor's (2016) 'technomoral practices': habits of ethical self-cultivation specifically mediated by technology. Lived Expressions make this tradition design-operational: they are mappable and, in principle, observable

features which might be changed by interaction with a given AI system, thus indicating that the Human Core has been affected – to the better or the worse.

While Lived Expressions are observable, they are not the center of our attention. Changes in Lived Expressions are, first of all, qualitative changes: different thoughts, different feelings, different behavior. For enabling pro-human design, such changes need to be assessed in terms of whether they indicate an improvement or a reduction of the dimensions of the Human Core. Sometimes this might be easy: Suppose that a person gets used to delegate important personal decisions (such as whether to split up with their partner) to an AI system because they think the AI system “understands me and my relationship better than I do”. This clearly indicates a loss of freedom and agency. Sometimes this might not be so clear: If a juvenile starts challenging his parents because of habitual conversations with an AI buddy telling him that his parents are oppressors, is this then an increase of autonomy (positive, because supporting personal development in this stage of life) or a decrease in relational capacity (thus negative), or both?

Modern psychology has developed models to infer from observables (behavior, but also thoughts and feelings, measured via specific test questions) to unobservable concepts such as happiness, resilience, relational capacity, etc., and it might be necessary to include this kind of knowledge (see Borsboom et al., 2003; Borsboom, 2005). In addition, different people might have different opinions of whether a change is positive or negative with respect to a specific dimension of the Human Core, depending on their ethical position. It is beyond the scope of this paper to work these elements out; it may suffice to state that this assessment is necessary, and that it includes an ethical element. This should not come as a surprise, as aligning technology with values is always an ethical endeavor, and it is well known from more classical areas of Responsible AI that design decisions usually involve ethical and normative aspects (see, e.g., Hertweck et al, 2023, for the case of Algorithmic Fairness). In the following, we assume that there is an assessment function which maps changes of Lived Expressions to a scale denoting “better” and “worse,” however always with respect to the earlier defined dimensions of the Human Core which determine the aspects of the humaneness we focus on.

The key analytical questions of Step 2 are therefore: (a) which Lived Expressions are affected by the given AI system in the given use case, and how so? And (b) are these changes positive or negative with respect to how well we can live out our humaneness according to the dimensions of the Human Core defined earlier.

Effects on the Human Core are assumed to be gradual and diffuse rather than immediate and dramatic: In most cases, the decisive mechanism is habitual repetition rather than single interactions with an AI system. This complicates the concrete measurement. From a practicality perspective, our methodology therefore focuses on the one or two characteristics most strongly affected by a given Base AI System's design, rather than cataloguing every marginal effect. This makes the approach tractable and the resulting design

interventions targeted. The analysis can be iterated once the efficacy of an initial intervention has been established.

### **3.3 Step 3: Designing Pro-human Interventions**

The third step translates the insights from Step 2 into a concrete design intervention that modifies the Base AI System toward a pro-human orientation (henceforth: the Pro-human AI System). The aim is to preserve, and where possible improve the condition of the affected characteristics of the Human Core, while retaining the (efficiency) gains that motivated the introduction of the AI system in the first place.

Step 3 then is a creative design act with the goal to preserve the identified Lived Expressions while retaining the AI system's utility. There is no universal template for how to do so: the relevant Lived Expressions and their relationship to the system's design differ fundamentally across contexts, making each instantiation of Step 3 an original design act. This situates pro-human AI design within the tradition of design science: following Simon (1996), the creation of purposeful artifacts is a rigorous form of inquiry in its own right; Hevner et al. (2004) operationalise this for information systems research, establishing that a design science contribution is evaluated not by whether it followed a prescribed procedure but by whether the artifact achieves its intended purpose in context. Rather than specifying a theoretical procedure for this step in the abstract, we demonstrate it through a concrete case in Section 4 and present preliminary results.

## **4. Results**

We applied the three-step methodology to a use case in psychiatric healthcare: psychiatric session reporting. This use case was selected because summarizing conversations is an application where AI systems are increasingly used, but at the same time psychiatric documentation goes beyond purely administrative recording: it constitutes a recurring professional practice (Lived Expression) in which reflective, interpretative, and relational activities are structurally embedded in what appears on the surface to be routine clinical documentation.

### **4.1 Identifying the Human Core Characteristics (Step 1)**

Step 1 from Section 3 applies directly: the five constitutive characteristics identified in Section 3.1 serve as the analytical lens for the use case below.

### **4.2 Examining Interactions in Psychiatric Session Reporting (Step 2)**

At first sight, psychiatric session reporting appears as the prototypical documentation task to automate: It consists of a concise summary of the medical history, the held session, and derived diagnoses as well as next steps. Based on an automated transcript from the recorded consultation and optionally the physician's notes, an LLM-based Base AI System

could be conceived as a useful tool for creating a decent draft, to be signed off by the physician with a minimum of manual rework. It would free scarce time resources of a person highly trained to work with people by having them spend less time with paperwork.

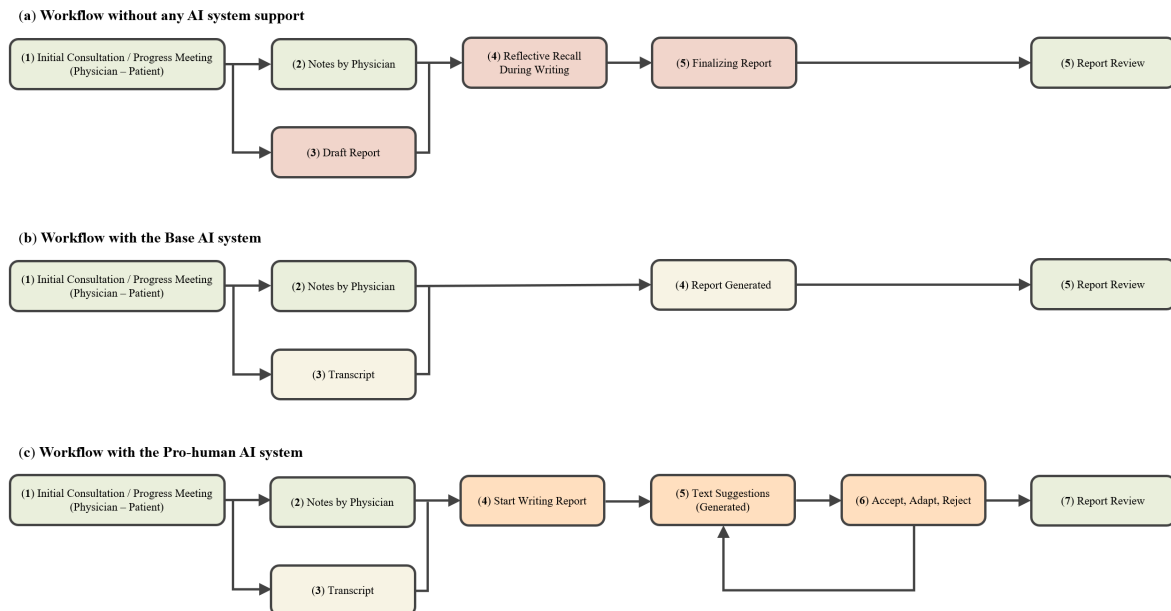


Figure 2. Comparison of psychiatric session reporting across three documentation processes: (a) workflow without an AI system, (b) workflow with the Base AI System, and (c) workflow with the Pro-human AI System that includes our design interventions based on the analysis of Human Core interactions.

However, the physician is tasked here not merely with an administrative burden. Instead, writing such a report serves not only for documentation purposes but is also a catalyst for reflection on the psychiatric session and thus constitutes a crucial step in the treatment process itself. In specific sections of the report, such as the mental status and the psychopathological findings, physicians do not merely summarize what was said and state simple observations, but, for example, reflect on how it was said, and the congruence (or absence thereof) between verbal content and nonverbal expression. They think about what should be reported and what should not, based on what might be important for the future or would be a mere distraction. They make conscious decisions on diagnosis fragments and also what aspects to keep deliberately open. Performing these reflections deliberately is a major factor in their felt responsibility for the patient and constitutive of the physician-patient relationship as well as for the diagnosis and next steps. In this sense, working on the report represents Lived Expressions (see Section 3.2) of connectedness (relating properly to the patient), freedom (forming conclusions without external determination), and agency (professional authorship of the therapeutic process).

When a Base AI System is introduced for psychiatric session reporting as discussed above, the documentation process changes fundamentally: The physician's role shifts from being the primary author to being an evaluator of a pre-formulated text (which itself has long-term detrimental effects on human involvement, see Stadelmann et al., 2026). Figure 2

.....

illustrates the three workflows: (a) without AI, (b) with the Base AI System, and (c) with the Pro-human AI System.

Based on this analysis, and corroborated by preliminary survey results, we predict that a Base AI System has noticeable detrimental effects on connectedness, freedom, and agency: their realization is structurally linked to Lived Expressions in the documentation process that would be streamlined away. Critically, this impact is assumed to arise not from one-off use but from repeated and increasingly habitual interaction with Lived Expressions. When a cognitively and professionally demanding practice is repeatedly off-loaded to an automated system, the neural and professional pathways sustaining that practice gradually atrophy, eroding not only skill but the motivational and relational dispositions tied to it (Lawrence, 2024; Dehaene, 2021). Over time, clinical intuition may be diminished, and the physician may become gradually alienated from their patient.

### **4.3 Pro-human Design Intervention: Assisted Text Completion (Step 3)**

Based on the analysis in Step 2, the documentation process was redesigned to remain a place of active wording and judgment formation. The goal is to have the human in the lead, not just in the loop, while still preserving efficiency gains. The key design intervention is a shift from automated report generation to assisted text completion, inspired by code completion used in software engineering (see Robbes & Lanza, 2010; Husein et al., 2025).

In the Pro-human AI System, the physician initiates and structures the report based on the patient consultation, their own notes, and reflective recollection. During the writing process, the system analyses the transcript, any existing notes, and the physician's ongoing text input and suggests only preliminary text segments (ghost text) without generating a complete report. Thoughts (sentences) start with the human; the system assists with details. The physician iteratively revises, accepts, or rejects these suggestions, remaining fully in charge of the content and formulation.

Figure 3 shows a screenshot from a preliminary implementation. The Pro-human AI System handles administrative segments such as family history, current medication, or living situation automatically; the physician enters keyword-based assessments and receives provisional text completions inline that require explicit interaction.

A preliminary survey of physicians using the Pro-human AI System confirms that both goals are achieved (human in the lead, efficiency preserved): documentation time is reduced considerably compared to the workflow without AI System, and physicians report a heightened sense of professional safety: a confidence that their clinical judgment and relational skills remain active and intact rather than displaced. This combination, efficiency without self-alienation, is precisely the goal pro-human AI design sets out to reach.

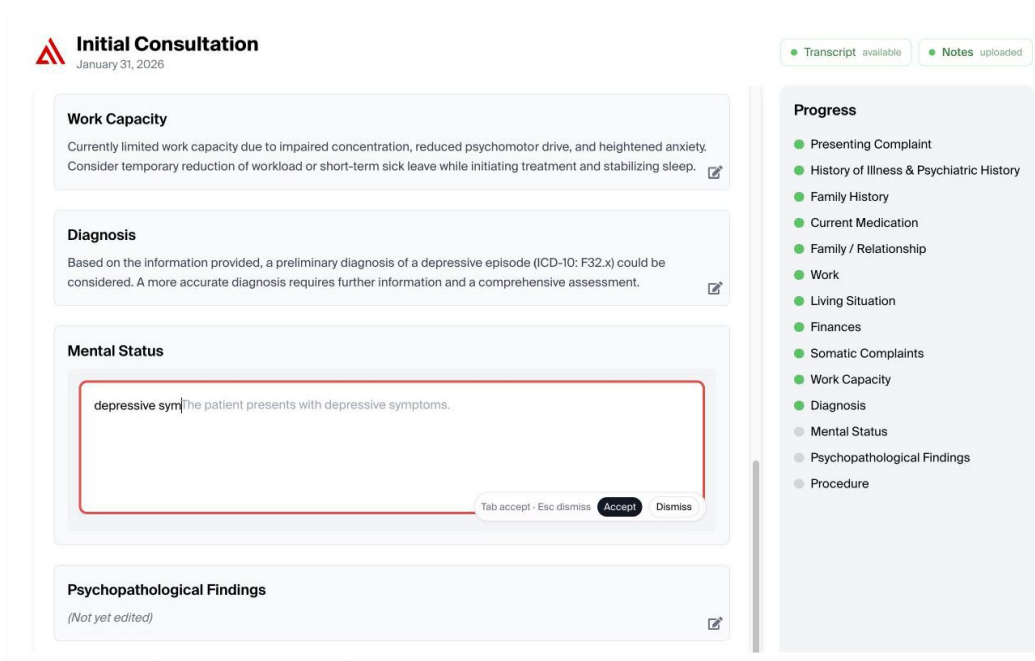


Figure 3. After entering keywords based on their assessment, the physician receives text suggestions from the Pro-human AI System, which they can either accept, modify, or reject through simple interactions (from von Wartburg-Kottler, 2026).

## 5. Discussion and Conclusions

### 5.1 Summary

This paper has introduced *pro-human AI design*: a three-step methodology for building AI systems that preserve, and ideally strengthen, the constitutive characteristics of being human. Grounding the *Human Core* in five characteristics (connectedness, freedom, agency, embodiment, and transcendence), the methodology examines how AI use engages their *Lived Expressions* in a given context and develops targeted technical interventions.

Applied to psychiatric session reporting, it operated on the theoretically motivated premise that the decisive risk lies not in AI use as such but in its habitual recurrence: when documentation is routinely automated, the Lived Expressions of professional authorship, reflective recall, and relational attentiveness are systematically displaced. A shift to assisted text completion, keeping the physician as primary author while offloading administrative detail, addresses this risk at least partly: preliminary results suggest that documentation time is reduced comparably to full automation, while physicians report a heightened sense of professional safety.

This approach does not represent a ready-to-use solution but a design direction, aiming to reduce the systematic weakening of the Human Core by maintaining the conditions for the practice of their Lived Expressions. Pro-human AI design is most urgently needed where AI use is recurring, relational, and cognitively formative, with dialogue systems in

therapeutic and counseling contexts as a prototypical case. That it can be done, that it can preserve efficiency, and that the methodology is applicable across domains is the hopeful contribution of this work.

## 5.2 Limitations

The theoretical architecture of the Human Core warrants further scrutiny. The five characteristics do not all sit at the same level of abstraction: embodiment is an ontological condition that underlies the others, freedom and agency are closely related faces of self-determination, and transcendence presupposes the freedom to question. A fuller analysis of these internal relationships, of whether the list achieves the strict minimality we claim for it, and whether the five characteristics' state can be automatically assessed or measured in a useful way, and remains future work.

The current analysis also rests on a single use case and a preliminary qualitative methodology. The potential effects on the Human Core characteristics are plausible inferences from the structure of the use case but were not empirically validated. Several central claims, such as about skill atrophy through habitual off-loading, about the relational effects of automated documentation, and about the efficacy of assisted text completion, are supported by theoretical argument and preliminary data, but await longitudinal empirical study.

A structural limitation runs deeper: even well-designed individual products cannot fully protect users who interact daily with different AI systems, not all of their choice, over which they have no design influence. This points toward an additional research and product innovation direction: a pro-human *adaptor layer* sitting between the user and any AI endpoint, one that maintains the conditions for the Lived Expressions of the Human Core regardless of the underlying model. Beyond benefiting individual users, such a layer would make the pro-human approach accessible to a broader base of potential suppliers, not just organizations with existing AI-based products.

## 5.3 A Call to Action: Join the Pro-Human AI Movement

We see pro-human AI design not merely as a research contribution but as a call to action for developers, designers, researchers, product managers, and business leaders alike. There are three reasons to join. (1) *we can*. The methodology is executable, the vocabulary is in place, and preliminary evidence is encouraging. The question is not whether pro-human AI design is possible but whether we will commit to doing it. (2) *we should*. AI systems deployed at scale do not merely process tasks but shape the habits and, ultimately, the characteristics of the humans who use them. The atrophy mechanism is real (Dehaene, 2021; Lawrence, 2024), and its consequences are already emerging (Fang et al., 2025). Designing systems that erode professional competence, relational attentiveness, or reflective judgment is not a neutral technical choice but a moral one. (3) *it is, in all likelihood, a good business idea*. In a market increasingly shaped by concerns about AI safety and societal impact, a demonstrably pro-human approach can become a powerful differentiator; in this

niche, a trust economy might outcompete an attention economy (see Stadelmann, 2026b). Ethical credibility combined with genuine and competitive user benefit are advantages that purely efficiency-driven systems cannot match.

#### **5.4 Future Work: Towards the HUMANCORE Benchmark**

As future work, we envision the HUMANCORE benchmark: a dedicated evaluation framework to measure the pro-human orientation of AI systems by their effects on the Human Core. Unlike performance-focused benchmarks and operationalizing the aspiration of initiatives like the Pro-Social AI Index (Walther, 2025), it would assess how a system’s design affects the conditions under which users can practice and develop their constitutive human characteristics, serving as a guide for design decisions and an incentive for developers to take the pro-human dimension seriously (amplified by its rate of adoption).

Designing such a benchmark will require departing from standard evaluation paradigms. Kommers et al. (2026), drawing on hermeneutic theory, argue that benchmarks for AI systems in cultural and human contexts should be iterative, involve human participants, and measure the contextual interaction between system and user rather than isolated model outputs. These principles align precisely with what a HUMANCORE benchmark must do: pro-human effects only emerge through habitual use, cannot be read off technical proxies, and are always context-specific. They can, however, be assessed relatively, making such a benchmark possible in ‘LLM-Arena style,’ but with the human as the ‘system under test’ of psychometric measures designed according to empirical psychological principles.

Pro-human AI design provides a structured methodology for asking a question that is too often overlooked: not just "What does this AI system do, and which harm can be caused by its actions?" but more specifically "What impact does this AI system have on the people who (habitually) use it?" While the first question is focusing on AI systems, the question that we propose here puts the human being in the center.

This is, ultimately, a question about safety: not of the human race from superintelligent systems, but of the individual user from the gradual erosion of what makes them fully human. This safety matters today, deserving a central place in the design of every AI system.

#### **Acknowledgements**

We are grateful for the transdisciplinary perspectives offered in feedback and discussions by Jan Segessenmann, Irina Raicu, Carmody Grey, Oliver Dürr, Johannes Hartl, Anthony King, Brittany Ehemann, Markus Christen, Birte Platow, Zoltan Schwab, and Stefan Huber. The result, including any errors, is ours.

#### **Literature**

Blattner, M. (2026). The Cognitive Resonance Framework: Aligning Generative AI with Durable Learning Through Staged Assistance. (*Under review*).

- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110(2), 203–219. <https://doi.org/10.1037/0033-295X.110.2.203>
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511490026>
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Buber, M. (1923). *Ich und Du*. Insel-Verlag.
- Burwell, J. M. (2025). The AI Efficiency Paradox: Reclaiming Quality Patient Care in an Era of Optimization. *Journal of Medical Systems*, 49(1), 49. <https://doi.org/10.1007/s10916-025-02183-2>
- Calo, Z. R. (2024). AI, medicine and Christian ethics. In B. Solaiman & I. G. Cohen (Eds.), *Research Handbook on Health, AI and the Law* (pp. 219–233). Edward Elgar Publishing. <https://doi.org/10.4337/9781802205657.ch13>
- Cassirer, E. A. (1960). *Was Ist der Mensch?: Versuch Einer Philosophie der Menschlichen Kultur*. W. Kohlhammer.
- Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), eaec8352. <https://doi.org/10.1126/science.aec8352>
- Comer, J. M. (2015). *Garden City: Work, Rest, and the Art of Being Human*. Grand Rapids: Zondervan.
- Crawford, M. B. (2015). *The world beyond your head: on becoming an individual in an age of distraction* (First edition ed.). New York: Farrar, Straus and Giroux.
- Dehaene, S. (2021). *How we learn: why brains learn better than any machine ... for now*. New York: Penguin Books.
- European Commission. (2024). *Artificial Intelligence Act*. <https://artificial-intelligence-act.eu>
- European Commission High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. *European Commission*. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). *How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study*. arXiv. <https://arxiv.org/abs/2503.17473>
- Fuchs, T. (2024). *Verteidigung des Menschen: Grundfragen einer verkörperten Anthropologie* (5. Auflage ed.) (No. 2311). Berlin: Suhrkamp.
- Fuchs, T., Aszmann, O., & Dürr, O. (2024). Organisms, prostheses and the limits of cyborgization. *Philosophy, Theology and the Sciences*, 11(2), 208–226. <https://doi.org/10.1628/ptsc-2024-0016>
- Gastmans, C., Sinibaldi, E., Lerner, R., Yáñez, M., Kovács, L., Palazzani, L., Pegoraro, R., & Vandemeulebroucke, T. (2024, November). Christian anthropology-based contributions to the ethics of socially assistive robots in care for older adults. *Bioethics*, 38(9), 787–795. <https://doi.org/10.1111/bioe.13322>
- Gehlen, A., & Rehberg, K.-S. (2016). *Der Mensch: seine Natur und seine Stellung in der Welt* (No. 89). Frankfurt am Main: Vittorio Klostermann.
- Google. (2018, June 7). *Artificial intelligence at Google: Our principles*. <https://blog.google/technology/ai/ai-principles/>
- Gunkel, D. J., & Wales, J. J. (2021). Debate: what is personhood in the age of AI? *AI & Society*, 36(2), 473–486. <https://doi.org/10.1007/s00146-020-01129-1>
- Grey, C. (2025). *Practices of communion: The task of an integral ecology*. Otheo.
- Hansen, M. (2024). From attention economy to cognitive lock-ins. *Big Data & Society*, 11(3). <https://doi.org/10.1177/20539517241275878>
- Harris, T., & Raskin, A. (2023, March). The AI dilemma [Podcast episode]. In *Your Undivided Attention*. Center for Humane Technology. <https://www.humanetech.com/podcast/the-ai-dilemma>

- Hertweck, C., Baumann, J., Loi, M., Viganò, E., Heitz, C. (2023): A Justice-Based Framework for the Analysis of Algorithmic Fairness-Utility Trade-Offs, <https://arxiv.org/abs/2206.02891v3>
- Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105.
- Hill, K., & Valentino-DeVries, J. (2025). What OpenAI Did When ChatGPT Users Lost Touch With Reality. *The New York Times*. <https://www.nytimes.com/2025/11/23/technology/openai-chatgpt-users-risks.html>
- Hirschauer, S. (2025). Tiere, Götter, Dinge, Tote und andere Aliens: Eine vergleichende Kartierung der Distinktionszonen des Humanen. In S. Hirschauer, P. Hofmann, A. Friedrichs, & G. Schabacher (Eds.), *Humandifferenzierung im Vergleich* (pp. 357–383). Velbrück. [10.5771/9783748962809-357](https://doi.org/10.5771/9783748962809-357)
- Holstein, K., Alevén, V., & Rummel, N. (2020). A conceptual framework for human–AI hybrid adaptivity in education. In I. Bittencourt et al. (Eds.), *Artificial Intelligence in Education. AIED 2020* (LNAI 12163, pp. 240–254). Springer. [https://doi.org/10.1007/978-3-030-52237-7\\_20](https://doi.org/10.1007/978-3-030-52237-7_20)
- Huber, S. (2026). »God« is still encountered – even by Reformed Christians in Switzerland. *Jahrbuch für Seelsorge, Spiritual Care Und Pastoralpsychologie*, 1, 25–45. <https://doi.org/10.36950/jssp.2026.1.3>
- Husein, R. A., Aburajouh, H., & Catal, C. (2025). Large language models for code completion: A systematic literature review. *Computer Standards & Interfaces*, 92, 103917.
- Ibrahim, L., Huang, S., Bhatt, U., Ahmad, L., & Anderljung, M. (2025). Towards interactive evaluations for interaction harms in human-AI systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Vol. 8, No. 2, pp. 1302–1310). AAAI Press. <https://ojs.aaai.org/index.php/AIES/article/view/36631>
- IEEE. (2021). IEEE Std 24748-7000-2021: IEEE standard model process for addressing ethical concerns during system design. *IEEE*. <https://doi.org/10.1109/IEEESTD.2021.9536679>
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (2nd ed.). *IEEE Standards Association*. <https://standards.ieee.org/industry-connections/ec/autonomous-systems/ead2e/>
- Ihde, D. (Ed.). (2010). *Technology and the lifeworld: from garden to earth*. Bloomington: Indiana University Press.
- Keane, W. (2024). *Animals, robots, gods: Adventures in the moral imagination*. Allen Lane.
- Kommers, C., Ahnert, R., Antoniak, M., Benetos, E., Benford, S., Bunz, M., Caramiaux, B., Concannon, S., Disley, M., Dobson, J., Du, Y., Duéñez-Guzmán, E., Francksen, K., Gius, E., Gray, J. W. Y., Heuser, R., Immel, S., Leigh, S., Livingston, D., Long, H., Martin, M., Meyer, G., Mihai, D., Noel-Hirst, A., Ostherr, K., Parker, D., Qin, Y., Ratcliff, J., Robinson, E., Rodriguez, K., Sobey, A., Underwood, T., Vashistha, A., Wilkens, M., Wu, Y., Zheng, Y., & Hemment, D. (2026). Computational hermeneutics: Evaluating generative AI as a cultural technology. *Frontiers in Artificial Intelligence*, 9, 1753041. <https://doi.org/10.3389/frai.2026.1753041>
- Lawrence, N. (2024). *The Atomic Human: Understanding Ourselves in the Age of AI*. Allen Lane.
- MacIntyre, A. (1981). *After virtue: A study in moral theory*. University of Notre Dame Press.
- Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Gallimard.
- Microsoft. (2022). Responsible AI principles. *Microsoft*. <https://www.microsoft.com/en-us/ai/principles-and-approach>
- OECD. (2019). OECD principles on AI. *OECD*. <https://oecd.ai/en/ai-principles>
- Plessner, H. (1928). *Die Stufen des Organischen und der Mensch: Einleitung in die philosophische Anthropologie*. Berlin Leipzig: Walter de Gruyter. <https://doi.org/10.1515/9783111537429>
- Ricoeur, P. (1992). *Oneself as another* (K. Blamey, Trans.). University of Chicago Press.
- Robbes, R., & Lanza, M. (2010). Improving code completion with program history. *Automated Software Engineering*, 17(2), 181–212.

- .....
- Rubin, M., Arnon, H., Huppert, J. D., & Perry, A. (2024). Considering the Role of Human Empathy in AI-Driven Therapy. *JMIR Mental Health*, 11, e56529. <https://doi.org/10.2196/56529>
- Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
- Scheler, M. (2016). *Die Stellung des Menschen im Kosmos*. S.I.: Eisenbrauns.
- Schirch, L., Slachmujlder, L., & Iyer, R. (2023). Toward Prosocial Tech Design Governance. *Tech Policy Press*, 14 December 2023. <https://www.techpolicy.press/toward-prosocial-tech-design-governance/>
- Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). MIT Press.
- Spaemann, R. (2006). *Persons: The difference between "someone" and "something"* (O. O'Donovan, Trans.). Oxford University Press.
- Spiekermann, S. (2019). *Ethical IT innovation: A value-based system design approach*. CRC Press.
- Stadelmann, T. (2025). A guide to AI: Understanding the technology, applying it successfully, and shaping a positive future. *Global Resilience White Papers*, No. 2. <https://www.globalresiliencepub.com/>
- Stadelmann, T. (2026a). Debate: Evidence-based AI risk assessment for public policy. *Public Money & Management*, 46(1). <https://doi.org/10.1080/09540962.2025.2541304>
- Stadelmann, T. (2026b, February 9). AI in 2035 - A hope-filled vision for a humane future with AI. *AIssais blog* (Original work published October 27, 2025). <https://stdm.github.io/AI-in-2035/>
- Stadelmann, T., Merkt, P. H., & Barr, K. (2026). The stochastic nature of machine learning and its implications for high-consequence AI. *AI and Ethics*, 6, 195. Springer. <https://doi.org/10.1007/s43681-026-01042-1>
- Segessenmann, J., Stadelmann, T., Davison, A., & Dürr, O. (2025). Assessing deep learning: a work program for the humanities in the age of artificial intelligence. *AI and Ethics*, 5(1), 1–32. <https://doi.org/10.1007/s43681-023-00408-z>
- Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press. <https://doi.org/10.1093/oso/9780192845290.001.0001>
- Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780190498511.001.0001>
- Vallor, S. (2024a). *The AI mirror: how to reclaim our humanity in an age of machine thinking*. New York (N.Y.): Oxford University Press.
- Vallor, S. (2024b). The Thoughts the Civilized Keep. In *The AI Mirror* (1st ed., pp. 102–132). Oxford University Press. <https://doi.org/10.1093/oso/9780197759066.003.0005>
- Verbeek, P.-P. (2005). *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Penn State University Press.
- Viganò, E., Hertweck, C., Heitz, C., & Loi, M. (2022). People are not coins: Morally distinct types of predictions necessitate different fairness constraints. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2293–2301). ACM. <https://doi.org/10.1145/3531146.3534643>
- van der Rijt, J.-W., Coelho Mollo, D., & Vaassen, B. (2026). AI mimicry and human dignity: Chatbot use as a violation of self-respect. *Journal of Applied Philosophy*. <https://doi.org/10.1111/japp.70037>
- Walther, C.C. (2024). ProSocial artificial intelligence as a catalyst for holistic health: a multidimensional approach. *BMC Global Public Health* 2, 76. <https://doi.org/10.1186/s44263-024-00111-z>
- Walther, C.C. (2025, August 30). A prosocial AI index: Measuring what matters for people and planet. *Forbes*. <https://www.forbes.com/sites/corneliawalther/2025/08/30/a-prosocial-ai-index-measuring-what-matters-for-people-and-planet/>
- Wehrli, S., Hertweck, C., Amirian, M., Glüge, S., & Stadelmann, T. (2021). Bias, awareness and ignorance in deep-learning-based face recognition. *AI and Ethics*. <https://doi.org/10.1007/s43681-021-00108-6>
- Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). *The rise of AI companions: How human-chatbot relationships influence well-being*. arXiv. <https://arxiv.org/abs/2506.12605>