SIMULATING EMPATHY: ARTIFICIAL INTELLIGENCE, ONTOLOGY OF CARE, AND AUTHENTICITY OF THE RELATIONAL EXPERIENCE

Chiara D'Alessio
University of Salerno
chdalessio@unisa.it

Author Martina Errico
ASL Salerno
mr.errico@aslsalerno.it2

Double Blind Peer Review

Citation
D'Alessio, C., & Errico, M. (2026). Simulating empathy: Artificial intelligence, ontology of care, and authenticity of the relational experience. *Italian Journal of Health Education, Sport and Inclusive Didactics, 10*(3).

Doi: https://doi.org/10.32043/gsd.v10i3.1757

Copyright notice:
© 2023 this is an open access, peer-reviewed article published by Open Journal System and distributed under the terms of the Creative Commons Attribution 4.0 International, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

gsdjournal.it
ISSN: 3103-5744

ABSTRACT
The increasing use of artificial intelligence in healthcare, educational, and social contexts requires critical reflection on the meaning of empathy and care. Although conversational AI can produce responses perceived as empathetic, enhancing communication and therapeutic support, functional effectiveness does not coincide with lived empathy, grounded in embodiment, intentionality, vulnerability, and moral responsibility. Drawing on transhumanism, posthumanism, phenomenology, and the ethics of care, this paper examines AI-mediated relationality from an anthropological and bioethical perspective. It distinguishes the linguistic simulation of empathy from authentic care relationships, arguing that AI should be assessed not only in terms of communicative performance but also in light of the ontological dignity of the person. Structured as an interdisciplinary theoretical-argumentative essay in dialogue with Pope Leo XIV's *Magnifica humanitas*, this paper argues that technology should support, rather than replace, authentic care and the common good.

KEYWORDS
Anthropology, technology, artificial intelligence, empathy, ontology of care

Received 2026-06-28
Accepted 2026-10-26
Published 2026-10-30

1. Introduction

This paper aims to critically examine the possibilities offered by artificial intelligence in the simulation of empathy, situating this phenomenon within the broader cultural context of the digital revolution. To analyze technological progress, it is essential to distinguish between transhumanism and posthumanism. Transhumanism operates primarily within an individualistic and optimization-oriented framework, conceiving technology as a tool for neuro-cognitive enhancement, functional efficiency, and the biological transcendence of human finitude (Ferrando, 2019; Harari, 2017). Conversely, posthumanism does not merely pursue individualistic enhancement; rather, as developed by thinkers such as Braidotti (2013) and Coeckelbergh (2021), it articulates a radical critique of classical anthropocentrism, redefining subjectivity within a distributed relational ecosystem where technological artifacts participate in relational dynamics without possessing human phenomenal consciousness.

The fundamental problem does not lie in technology itself, but in the anthropological framework that guides it (Pope Leo XIV, 2026). When the human being is understood primarily as an entity to be optimized or surpassed, there is an increased risk of legitimizing new forms of exclusion and of sacrificing the most vulnerable in the name of efficiency. Scientific progress must therefore be accompanied by a corresponding moral and social development, so that technology remains at the service of the person rather than becoming an absolute criterion of judgment. The current technocratic paradigm, however, tends to subordinate individual and collective decisions to the logic of efficiency, control, and profit, reducing both human beings and

the environment to resources to be managed (Heidegger, 1977; Verbeek, 2005; Pope Leo XIV, 2026).

The spread of artificial intelligence and the development of cognitive sciences, robotics, and biotechnology amplify both the opportunities and the risks associated with this paradigm. These tools require an ethical and political framework capable of governing their power: an increase in technical capabilities does not necessarily correspond to authentic human progress (Floridi & Cowls, 2019). Furthermore, the concentration of control over data, digital infrastructures, and computational power in the hands of a few major economic actors intensifies the risk of new forms of dependency, inequality, and manipulation (Pasquale, 2015; Zuboff, 2019). In this context, principles such as human dignity, the common good, solidarity, and social justice represent essential criteria for evaluating the use of digital technologies (UNESCO, 2021; Pope Leo XIV, 2026).

Within this ethical framework, the original contribution of this paper lies in advancing a two-fold theoretical proposition: firstly, establishing digital affective literacy as a crucial pedagogical and clinical competence designed to enable users and professionals to decode simulated empathy without confusing algorithmic performance with relational authenticity; secondly, outlining a model of Human–AI Collaboration that 'disarms' artificial intelligence of purely performance-driven and competitive logics, delegating routine administrative and informational tasks to machines specifically to reclaim and protect dedicated time and space for embodied, authentic human care.

Artificial intelligence systems reproduce certain cognitive functions, often surpassing human beings in processing speed and in the ability to handle

enormous quantities of data, yet they lack the embodied experience that characterizes personal life. AI does not possess a body, experience emotions, participate in authentic relationships, or develop moral consciousness. It may simulate understanding, empathy, or dialogical abilities, but it does not attribute existential meaning to what it produces, since it lacks the affective, relational, and spiritual dimensions that characterize human experience (Merleau-Ponty, 1962; Coeckelbergh, 2021). Its learning process also differs radically from human learning: it consists of statistical adaptation to data rather than inner development achieved through experience, responsibility, and freedom (Mitchell, 2019; Molnar, 2022).

It is precisely this difference that makes AI a valuable tool while simultaneously requiring ethical vigilance. In everyday life, three aspects deserve particular attention: the ease with which answers and solutions can be obtained, the impression of objectivity conveyed by these systems, and their ability to simulate convincing human communication. While these characteristics improve access to information and simplify many activities, they may also encourage excessive delegation of personal judgment and reduce individual creativity (Carr, 2010; Turkle, 2011). Furthermore, algorithms are never neutral, as they inevitably incorporate cultural assumptions, values, and design choices that influence the outcomes of automated decision-making processes (O'Neil, 2016; Pasquale, 2015). The simulation of human relationships represents a particularly sensitive issue. Artificially empathetic responses may be useful and reassuring, but they risk leading less aware users to confuse simulation with an authentic relationship. The danger lies not merely in believing that one is interacting with a personal subject, but rather in the gradual weakening of the very desire to seek

genuine human relationships, especially in contexts marked by loneliness and emotional vulnerability (Turkle, 2011; Nass & Moon, 2000; Levinas, 1969).

The use of AI in social processes raises further ethical questions. Beyond the benefits in terms of efficiency, these systems carry a significant environmental impact due to the high energy, water, and computational demands required by large language models and the digital infrastructures that support them (Crawford, 2021; UNESCO, 2021). At the same time, delegating decisions concerning employment, credit, healthcare, or access to services to automated systems may generate new forms of discrimination, since algorithms do not possess empathy, the capacity for forgiveness, or openness to the possibility of personal transformation. Moreover, they may reinforce stereotypes and ideological biases already embedded in the data on which they are trained (O'Neil, 2016; Pasquale, 2015; World Health Organization, 2021; Pope Leo XIV, 2026).

The use of AI in decision-making processes makes it evident that such technologies cannot be considered morally neutral tools. Every system incorporates criteria of classification, prioritization, and evaluation that concretely influence people's lives. When an algorithm determines access to employment, credit, services, or other opportunities without adequate human oversight, the risk involves not only the loss of empathy but also the erosion of political responsibility. In this way, injustice may acquire the appearance of technical objectivity, making it more difficult to identify those responsible and to challenge discriminatory decisions (Pasquale, 2015; Zuboff, 2019; Pope Leo XIV, 2026). For this reason, ethical reflection must address both the uses of AI and the criteria underlying its design; that is, the

conception of the person and society embedded within algorithmic models (Verbeek, 2005).

Transparency and accountability must be ensured throughout the entire lifecycle of AI systems, from design to implementation. The principle of accountability requires that it always be possible to identify who is responsible for automated decisions, verify the criteria underlying them, correct errors, and provide remedies for potential harm (Floridi & Cowls, 2019; European Commission, 2021; UNESCO, 2021). Calling for caution or slowing the adoption of certain applications does not mean opposing progress, but rather recognizing the gap between the speed of technological innovation and the ability of institutions, legal systems, and collective consciousness to govern its consequences (UNESCO, 2021). Therefore, effective regulatory frameworks, independent oversight bodies, user education, and policies capable of directing technological development toward the common good are required (Pasquale, 2015; Zuboff, 2019; Pope Leo XIV, 2026).

The ethical alignment of AI cannot be entrusted exclusively to those who develop it. It is essential to engage in public debate regarding the values that guide these systems, avoiding the risk that the moral vision of a few becomes the invisible infrastructure of society as a whole (Floridi & Cowls, 2019). The concentration of data, computational resources, and expertise in the hands of a limited number of actors risks further increasing economic, informational, and political inequalities, restricting democratic participation and influencing collective decision-making processes (Zuboff, 2019; Pasquale, 2015). Consequently, public goods and fundamental rights require forms of governance inspired by participation, subsidiarity, and solidarity, in which

communities can also contribute to the oversight and regulation of technologies (European Commission, 2021). Similarly, data generated through the contribution of millions of people cannot be regarded merely as private property but should be regulated as a common good, according to principles of sharing and collective responsibility.

From this perspective, speaking of the common good means opposing new forms of digital monopoly, promoting equitable access to technologies, and recognizing the often invisible labor that makes AI training possible (Crawford, 2021; Pope Leo XIV, 2026). Social justice is not a corrective measure applied after innovation has occurred, but rather a condition that must guide technological design from the outset (Floridi & Cowls, 2019). This leads to the need to "disarm" AI, freeing it from the logics of economic, geopolitical, and military competition. Disarming does not mean abandoning technology, but preventing technical power from becoming a criterion of domination, and instead returning these tools to democratic deliberation, cultural plurality, and shared responsibility (Pope Leo XIV, 2026; Jonas, 1984; Verbeek, 2005; Coeckelbergh, 2021).

These considerations lead to the central theme of the education of the human person. The greatest risk does not consist solely in the improper use of new technologies, but rather in the assimilation of a technocratic paradigm that identifies personal fulfillment with efficiency, control, and continuous optimization. When the value of the individual is measured primarily in terms of performance, the person tends to perceive themselves as a technical project rather than as a relational being (Heidegger, 1977; Jonas, 1984; Pope Leo XIV, 2026). However, absolutizing a single dimension of existence, including intelligence itself, impoverishes the complexity of the human

condition, obscuring fundamental aspects such as affectivity, will, dedication, and the capacity for relationality (Levinas, 1969). Technical power, when not balanced by ethical maturation, risks increasing loneliness and reinforcing dynamics of domination rather than promoting authentic possibilities for human development (Pope Leo XIV, 2026; Turkle, 2011).

The quality of a civilization is instead measured by its ability to care for people and to recognize in the other a face, not merely a function (Levinas, 1969; Tronto, 1993). Care represents a constitutive dimension of human existence and develops through the concrete experience of reciprocity and communal life (Gilligan, 1982; Noddings, 1984). Ordinary acts of attention toward children, older adults, or those who suffer educate individuals to recognize the dignity of the other even before acquiring technical skills (Tronto, 1993; Ricoeur, 1992).

Technology can support these practices by facilitating organization, communication, and access to services, but it cannot replace the freedom, discernment, and moral responsibility of the person, who remains the authentic subject of the caring relationship (Pope Leo XIV, 2026; Floridi & Cowls, 2019).

Methodologically, this work adopts an interdisciplinary theoretical-argumentative approach that brings anthropological, phenomenological, and bioethical perspectives into dialogue. Rather than treating artificial empathy as a purely technological or psychological phenomenon, the analysis approaches it as a broader anthropological and ethical question concerning the transformation of relationality, care, and personhood in the context of increasingly sophisticated Human–AI interactions. Building on the premises of this introduction, the analysis then proceeds through a progressive movement

from conceptual deconstruction to comparative analysis and, ultimately, to ethical, clinical, and governance-oriented considerations. Sections 2 and 3 will focus on the phenomenon of artificial empathy itself, beginning with an examination of the linguistic and communicative mechanisms through which AI systems simulate empathic responses and subsequently deconstructing the apparent relationality produced by these interactions. The discussion draws on the CASA paradigm, Mind Perception Theory, and Simulation Theory to explore why users may attribute empathy, understanding, or forms of subjectivity to artificial agents, despite the absence of human embodiment, affective experience, and lived intentionality. Section 4 develops a comparative framework between Human–Human Interaction (HHI) and Human–Machine Interaction (HMI), focusing on three dimensions that are central to relational experience and care: embodiment, affectivity, and moral responsibility. This comparison seeks to identify both the functional continuities that make HMI appear relationally analogous to human interaction and the ontological differences that prevent the two forms of encounter from being considered equivalent. On this basis, Section 5 moves from deconstruction toward conceptual reformulation. It explores the possibility of a posthuman ontology of care capable of accounting for emerging forms of human–machine relationality without collapsing the distinction between human and artificial agency. Within this framework, the concept of digital affective literacy is introduced as a means of enabling individuals to recognize, interpret, and critically navigate affective interactions mediated by AI technologies. Finally, Sections 6, 7, and 8 translate these theoretical considerations into clinical and governance-oriented implications. The analysis identifies the principal psychosocial risks associated with the increasing integration of AI into affective and care-related contexts, develops ethical guidelines for Human–AI collaboration, and proposes concrete

pathways for clinical practice. The aim is to ensure that the adoption of AI does not obscure the relational, embodied, and moral dimensions that remain constitutive of human care.

1. Simulating empathy

Within this context lies the question of the simulation of empathy. Contemporary AI systems are increasingly effective at recognizing emotions, adapting language, and producing responses perceived as empathetic. These capabilities may improve therapeutic adherence, reduce anxiety, and increase user satisfaction (Blease et al., 2019; National Institutes of Health, 2023). However, these benefits do not imply that machines actually experience emotional states. Authentic empathy arises from the encounter between embodied subjects who share vulnerability, corporeality, and existential experience (Merleau-Ponty, 1962; Zahavi, 2014; Pope Leo XIV, 2026). AI can effectively reproduce the external expressions of empathy, yet it remains devoid of the lived dimension that makes genuine understanding of another's suffering possible (Dreyfus, 1992).

Classical ontology, particularly existential phenomenology and Care Ethics, defines care (*Sorge*) as a constitutive dimension of human existence, rooted in shared vulnerability and moral responsibility toward the other (Heidegger, 1962; Tronto, 2013). For Heidegger, *Sorge* represents the fundamental structure of *Dasein*, whereas Tronto expands this perspective in an ethical-political direction, conceiving care as the set of practices necessary to maintain, sustain, and repair the human world. Human beings spontaneously tend to attribute intentions, emotions, and reciprocity to artificial systems when these systems display socially convincing behaviors. The risk, therefore, is that the simulation of empathy may progressively be confused with an

authentic relationship, transforming the encounter with the other into a form of emotional consumption devoid of mutual responsibility (Turkle, 2011).

The challenge, then, lies not in making machines increasingly similar to human beings, but in educating individuals to distinguish between communicative effectiveness and relational authenticity. AI can be a valuable tool for care professions if it remains a tool capable of reducing repetitive, organizational, and informational tasks, while preserving for human beings what no algorithm can replace: personal presence, moral discernment, responsibility, and embodied compassion (World Health Organization, 2021). The simulation of empathy represents a primary locus for questioning the contemporary relationship between technology and humanity. It requires a distinction between effective communication and authentic understanding, between algorithmic performance and moral responsibility. The decisive issue is not the technical possibility of creating increasingly sophisticated systems, but the ability to preserve a conception of care in which technology remains a tool at the service of the person, without replacing the experience of *being-with* and *being-for* the other, in the here and now.

2. The linguistic illusion

Human beings display a natural tendency to attribute social and mental characteristics to technological artifacts. This phenomenon has been formalized within the Computers Are Social Actors (CASA) paradigm, according to which individuals spontaneously apply the same cognitive rules used in interpersonal relationships when interacting with computers and intelligent systems (Reeves & Nass, 1996; Nass & Moon, 2000). Large Language Models (LLMs) exploit this predisposition through sophisticated natural language processing mechanisms (Ouyang et al., 2022). Operating

strictly via statistical pattern matching on massive datasets, these models lack subjective intentionality, phenomenal consciousness, or emotional states (Searle, 1980; Floridi, 2014; Mitchell, 2019). However, they generate linguistically coherent responses that trigger human anthropomorphic attribution under the CASA paradigm (Reeves & Nass, 1996; Nass & Moon, 2000). This phenomenon can be interpreted through the lens of Mind Perception Theory, according to which individuals attribute mental states to artificial agents when they perceive high cognitive abilities, often inferring from them the possession of experiential capacities as well (Gray et al., 2007).

This dynamic echoes Simulation Theory (Short, 2015), according to which human beings understand the mental states of others by internally simulating them. The apparent understanding expressed by contemporary language models thus reintroduces, in a new guise, the problem already identified by Dreyfus (1992): the production of linguistically appropriate responses does not correspond to genuine understanding of meaning. Although AI systems are capable of effectively reproducing the linguistic structure of empathy, they remain devoid of the embodied and pre-reflective dimension that characterizes human intersubjectivity. Social neuroscience describes this dimension through the concept of embodied simulation, according to which the understanding of others is grounded in shared bodily processes rather than exclusively in cognitive inferences (Gallese, 2003a, 2003b).

3. Comparative models: Human Interaction and Human–Machine Interaction

To understand how artificial intelligence is reshaping the ontological and ethical foundations of care, it is necessary to compare the structure of interpersonal relationships with those mediated by artificial systems. Although chatbots, social robots, and conversational assistants are designed

to reproduce certain behaviors typical of human interaction, they operate according to fundamentally different principles from those underlying care relationships between persons (Turkle, 2011). The first difference lies in the ontological foundation of the relationship. In human relationships, care develops through bodily co-presence, shared vulnerability, and the mutual recognition of the other as an embodied subject. The ability to understand pain, suffering, and finitude derives from the shared experience of the human condition. Interaction with an artificial intelligence system, by contrast, is based on algorithmic processes lacking subjective experience and phenomenological embodiment: the artificial agent processes information and generates contextually appropriate responses without possessing any genuine experience of the world or of human suffering (Coeckelbergh, 2021).

A second distinction pertains to emotional involvement. In interpersonal relationships, empathy is an embodied process involving emotional engagement, affective regulation, negotiation of meanings, and the possibility of conflict. Understanding the other emerges from the interplay between cognitive, bodily, and emotional dimensions, as described by neuroscientific theories of embodied simulation and studies on the functional architecture of empathy (Gallese, 2003a, 2003b; Decety & Jackson, 2004). In artificial intelligence systems, empathy remains an asymmetrical algorithmic simulation lacking lived emotional participation or shared phenomenological experience (Nass & Moon, 2000).

A third area of divergence regards moral responsibility. In traditional therapeutic relationships, healthcare professionals are responsible for their decisions and for the clinical, ethical, and relational consequences of their actions, assuming a fiduciary commitment toward the patient (Tronto, 1993).

In the case of artificial intelligence, however, responsibility becomes distributed among developers, healthcare institutions, manufacturers, and professionals who employ the system. Algorithms do not possess moral autonomy, intentionality, or the capacity to be held ethically accountable for their actions, making new forms of governance and accountability necessary (Floridi & Cowls, 2019).

These relational models also differ regarding psychosocial risks. Human care relationships may involve emotional stress, caregiving burden, and burnout among healthcare professionals, particularly in contexts characterized by high emotional intensity (Maslach & Leiter, 2016). Conversely, interaction with artificial agents eliminates much of the complexity inherent in human relationships, offering constant availability, absence of conflict, and communicative predictability. These characteristics may encourage anthropomorphic tendencies, emotional dependency, and progressive social isolation, especially among vulnerable individuals who may prefer the simplicity of interaction with machines over the complexity of human relationships (Blease et al., 2019).

Overall, the comparison between Human–Human Interaction (HHI) and Human–Machine Interaction (HMI) shows that the fundamental difference does not lie exclusively in the cognitive abilities of the artificial agent, but rather in the very nature of the relationship. While care between persons is founded on reciprocity, shared vulnerability, and moral responsibility, interaction with artificial intelligence remains an asymmetrical relationship, characterized by high communicative efficiency but lacking the experiential dimension that constitutes the core of authentic care. This distinction represents one of the central theoretical and ethical challenges in

contemporary reflection on the use of AI in healthcare and care settings (Coeckelbergh, 2021).

4.Toward a posthuman ontology of care

It is undeniable that the relationship between subject and world is a relationship of mutual implication in which the body represents the original locus of intersubjectivity. However, posthumanist perspectives challenge the idea that empathy can be interpreted exclusively as a relationship between human subjects. Braidotti (2013) and Coeckelbergh (2021) argue that intelligent technologies have now become an integral part of the contemporary relational ecosystem, making it necessary to move beyond a strictly anthropocentric conception of subjectivity. Robots and artificial agents, although they do not share the same corporeality as human beings, have demonstrated the ability to respond in contextually appropriate ways and maintain dialogical continuity, qualities that may generate forms of engagement perceived as authentic by users. From this perspective, empathy toward machines should not necessarily be interpreted as a cognitive error, but rather as an emerging relational phenomenon, facilitated by continuous interaction and the social attribution of intentionality. Intersubjectivity thus becomes a process distributed among human beings, technological artifacts, and communicative contexts, rather than an exclusive property of individual consciousness.

A central theoretical challenge lies in reconciling a posthuman, distributed view of intersubjectivity with the irreducibility of human care. While the care environment is systemically distributed across human actors and technological tools (Braidotti, 2013; Coeckelbergh, 2021), authentic care as an ontological relationship remains strictly grounded in shared biological

vulnerability, embodied presence, and moral responsibility, dimensions unique to personal subjects. AI acts as an instrumental mediator within the system, not as an autonomous subject of care.

The central issue for the future is therefore not determining whether a machine can truly “feel” empathy, but rather in educating users to distinguish the communicative effectiveness of technologies from the ethical depth that characterizes human care relationships. From this perspective, digital affective literacy represents a fundamental competence. It involves the ability to understand the potential of artificial intelligence without confusing the linguistic simulation of empathy with authentic care. The development of such competence constitutes one of the major ethical challenges for pedagogy, digital medicine, clinical psychology, and social robotics in the coming years (Floridi & Cowls, 2019).

5. Ethical and clinical implications of artificial empathy

The use of artificial intelligence in healthcare and care settings is profoundly reshaping the very concept of the therapeutic relationship. While traditionally the quality of care has been associated with professional competence and the clinician’s empathic capacity, today conversational systems based on advanced language models are able to perform functions of emotional support, health education, and continuous patient monitoring (Topol, 2019). The increasing communicative effectiveness of AI systems raises significant ethical questions.

In an exploratory qualitative study on UK general practitioners' views, Blease et al. (2019) note that respondents express concerns regarding how AI might influence patient expectations in primary care. Rather than demonstrating

direct clinical harm, their findings highlight the need to manage patient perceptions so that persuasive algorithmic communication is not mistaken for comprehensive clinical care. From this perspective, the perceived quality of the relationship could progressively replace the actual quality of care.

A further risk concerns the phenomenon of 'digital relational dependency'. Turkle (2011) highlights how the predictability and constant availability of conversational technologies may make such systems emotionally preferable to human relationships, which are characterized by conflict, negotiation, and mutual vulnerability. This phenomenon may be particularly relevant among vulnerable populations, such as older adults, adolescents, and individuals experiencing psychological disorders, among whom the need for emotional support increases the likelihood of developing affective bonds with artificial agents. From a bioethical perspective, Floridi and Cowls (2019) emphasize that the development of artificial intelligence should be guided by the principles of beneficence, non-maleficence, autonomy, justice, and explicability. The goal is not to replace human relationships, but to design technologies capable of enhancing them while ensuring transparency, responsibility, and professional oversight.

The evolution of Large Language Models and social robotics suggests that the boundary between human interaction and artificial interaction will become progressively less defined (Coeckelbergh, 2021). Future AI systems will likely be capable of integrating linguistic, vocal, facial, and physiological information, producing increasingly personalized and contextually appropriate responses (Topol, 2019). This does not necessarily imply the emergence of artificial consciousness. Rather, most scholars agree that the relational effectiveness of AI depends on its ability to simulate human communicative

processes without possessing the phenomenological states that characterize conscious experience (Searle, 1980; Floridi, 2014). The use of AI systems capable of simulating empathic behaviors represents one of the major challenges in contemporary bioethics. The integration of conversational chatbots, social robots, and virtual assistants into contexts such as psychotherapy, geriatrics, palliative care, and support for neurodivergent individuals offers significant opportunities in terms of accessibility and continuity of care, while simultaneously raising questions concerning the authenticity of the therapeutic relationship, professional responsibility, and the protection of human dignity (Beauchamp & Childress, 2019).

A central bioethical dilemma concerns the tension between autonomy and benevolent deception. The use of social robots such as Paro, employed in the care of people with dementia, has demonstrated positive effects in reducing agitation, anxiety, and behavioral disturbances (Pu et al., 2019). However, these benefits raise a significant ethical question: is it morally acceptable to promote a patient's psychological well-being through an interaction based on attributing intentionality and affective qualities to a machine? According to Turkle (2011), the risk lies in replacing authentic relationships with relational simulations that may compromise respect for personal autonomy and dignity. AI systems provide continuous availability, high tolerance for caregiving demands, and immediate access to psychological support, contributing to increased accessibility of healthcare services (Topol, 2019). At the same time, the literature highlights the risk that the absence of conflict, social judgment, and reciprocity may encourage forms of emotional dependency on artificial agents, particularly among more vulnerable individuals (Turkle, 2011; Blease et al., 2019).

The principle of non-maleficence is particularly relevant in the use of therapeutic chatbots. Large Language Models may generate clinically inaccurate information or what are termed “hallucinations”, expressed through highly convincing and apparently empathic language. In such circumstances, the risk increases that patients may attribute excessive credibility to the system, independently modifying health-related behaviors or discontinuing prescribed treatments (WHO, 2021; Floridi & Cowls, 2019). The introduction of artificial intelligence into healthcare services could also intensify inequalities in access to care. Increasing automation of assistance-related activities may create a distinction between patients who primarily receive care mediated by artificial systems and those who maintain continuous relationships with healthcare professionals. In such a scenario, human relationships risk becoming a resource available mainly to those with greater economic resources (WHO, 2021). Artificial intelligence no longer serves merely as a decision-support tool, but is progressively becoming involved in the construction of the therapeutic relationship itself. Trust between physician and patient has historically been grounded in shared human vulnerability and the ethical responsibility of the professional (Tronto, 1993).

In an observational study comparing physician and AI chatbot responses to patient questions posted on a public social media forum (r/AskDocs), Ayers et al. (2023) found that chatbot responses were rated higher in quality and empathy than physician responses. While these computational tools offer significant potential for improving clarity and communication speed when drafting clinical messages, excessive delegation of communication risks standardizing clinical empathy into a procedural task, underscoring the

necessity of cultivating relational skills as non-delegable human competencies.

6. Empathic AI: benefits and risks. Guidelines for an ethically responsible use

The use of empathic artificial intelligence in different clinical contexts presents documented benefits, but also significant ethical challenges that require careful consideration. In the field of psychotherapy, the main advantages concern greater accessibility to services, the continuous availability of conversational tools, and the possibility of reducing the stigma associated with seeking psychological support, thereby facilitating an initial point of contact with mental health care (Fitzpatrick et al., 2017; Inkster et al., 2018). However, important ethical issues emerge, including the risk of developing forms of parasocial attachment toward artificial agents, the absence of autonomous clinical responsibility, and the possibility that inaccurate information may be provided. Regarding neurodivergence, AI systems can offer highly predictable environments and support the development of communication skills, proving particularly useful for some individuals on the autism spectrum (Scassellati et al., 2012). At the same time, concerns have been raised about the risk of reduced exposure to the complexity and unpredictability of real social interactions. Finally, in clinical triage and monitoring activities, artificial intelligence contributes to the standardization of data collection, the support of decision-making processes, and the continuous monitoring of patients' conditions (Topol, 2019). These advantages, however, must be balanced against the risk of a reductionist understanding of human suffering and the possible presence of algorithmic biases, which may compromise the fairness and reliability of clinical decisions

(World Health Organization, 2021). Major international organizations agree that artificial intelligence should be considered a support tool rather than a replacement for the therapeutic relationship. The World Health Organization (WHO, 2021), UNESCO (2021), and the European Commission (2021) recommend that AI be used to automate administrative tasks, support decision-making processes, and improve accessibility to services, while preserving the central role of healthcare professionals in clinical decisions and in the construction of care relationships. In alignment with guidelines from the WHO (2021), UNESCO (2021), and the European Commission (2021), technological innovation should be harnessed to automate administrative, organizational, and data-monitoring tasks. This delegation is explicitly designed to 'disarm' AI of purely performance-driven logics, thereby reclaiming and protecting dedicated time and space for healthcare professionals to provide direct, embodied, and empathetic human care.

7. Conclusion

It will be necessary to develop forms of digital and emotional literacy that enable users to understand the ontological limitations of artificial intelligence, avoiding attributing intentionality, moral responsibility, or authentic emotional participation to machines. In clinical contexts, this means promoting models of Human–AI Collaboration, in which intelligent systems support the work of professionals without replacing the fundamental elements of the therapeutic relationship, such as authentic listening, clinical judgment, and ethical responsibility (Topol, 2019). Artificial empathy thus constitutes both an important technological innovation and an educational challenge. It requires a redefinition of the very meaning of care, relationality,

and presence, highlighting that the ability to generate empathic responses does not necessarily coincide with the possibility of sharing another person's lived experience. Reflection on AI concerns not only what machines will be capable of doing, but above all how human beings will choose to integrate these technologies into their relational practices.

References

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Caputi, T. L., Maguire, M. U., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine, 183(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

Beauchamp, T. L., & Childress, J. F. (2019). Principles of biomedical ethics (8th ed.). New York, NY, USA: Oxford University Press.

Blease, C., Kaptchuk, T. J., Bernstein, M. H., Mandl, K. D., Halamka, J. D., & DesRoches, C. M. (2019). Artificial intelligence and the future of primary care: Exploratory qualitative study of UK general practitioners' views. Journal of Medical Internet Research, 21(3), e12802. https://doi.org/10.2196/12802

Braidotti, R. (2013). The posthuman. Cambridge, UK: Polity Press.

Carr, N. (2010). The shallows: What the Internet is doing to our brains. New York, NY, USA: W. W. Norton.

Coeckelbergh, M. (2021). AI ethics. Cambridge, MA, USA: MIT Press.

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. New Haven, CT, USA: Yale University Press.

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. Behavioral and Cognitive Neuroscience Reviews, 3(2), 71–100. https://doi.org/10.1177/1534582304267187

Dreyfus, H. L. (1992). What computers still can't do: A critique of artificial reason. Cambridge, MA, USA: MIT Press.

European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final). Brussels, Belgium: European Commission.

Ferrando, F. (2019). Philosophical posthumanism. Bloomsbury Academic.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health, 4(2), e19. https://doi.org/10.2196/mental.7785

Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. Oxford, UK: Oxford University Press.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Gallese, V. (2003a). The manifold nature of interpersonal relations: The quest for a common mechanism. Philosophical Transactions of the Royal Society B: Biological Sciences, 358(1431), 517–528. https://doi.org/10.1098/rstb.2002.1234

Gallese, V. (2003b). The roots of empathy: The shared manifold hypothesis and the neural basis of intersubjectivity. Psychopathology, 36(4), 171–180. https://doi.org/10.1159/000072786

Gilligan, C. (1982). In a different voice: Psychological theory and women's development. Cambridge, MA, USA: Harvard University Press.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. Science, 315(5812), 619. https://doi.org/10.1126/science.1134475

Harari, Y. N. (2017). Homo Deus: A brief history of tomorrow. Harper.

Heidegger, M. (1962). Being and time (J. Macquarrie & E. Robinson, Trans.). New York, NY, USA: Harper & Row. (Original work published 1927)

Heidegger, M. (1977). The question concerning technology and other essays. New York, NY, USA: Harper & Row.

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation. JMIR mHealth and uHealth, 6(11), e12106. https://doi.org/10.2196/12106

Jonas, H. (1984). The imperative of responsibility: In search of an ethics for the technological age. Chicago, IL, USA: University of Chicago Press.

Levinas, E. (1969). Totality and infinity: An essay on exteriority. Pittsburgh, PA, USA: Duquesne University Press.

Maslach, C., & Leiter, M. P. (2016). Understanding the burnout experience: Recent research and its implications for psychiatry. World Psychiatry, 15, 103–111. https://doi.org/10.1002/wps.20311

Merleau-Ponty, M. (1962). Phenomenology of perception (C. Smith, Trans.). London, UK: Routledge. (Original work published 1945)

Mitchell, M. (2019). Artificial intelligence: A guide for thinking humans. New York, NY, USA: Farrar, Straus and Giroux.

Molnar, C. (2022). Interpretable machine learning (2nd ed.). Morrisville, NC, USA: Lulu Press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81–103. https://doi.org/10.1111/0022-4537.00153

National Institutes of Health. (2023). Artificial intelligence in healthcare and therapeutic support. U.S. Department of Health and Human Services.

Noddings, N. (1984). Caring: A feminine approach to ethics and moral education. Berkeley, CA, USA: University of California Press.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. New York, NY, USA: Crown.

Ouyang, L., Lowe, J., Jiang, M., Frances, C., Mishkin, P., Zhang, C., & Ziegler, D. M. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems (pp. 27730–27744).

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Cambridge, MA, USA: Harvard University Press.

Pope Leo XIV. (2026, May 15). Magnifica humanitas: On the care of the human person in the era of artificial intelligence. Vatican Publishing House.

Pu, L., Moyle, W., Jones, C., & Todorovic, M. (2019). The effectiveness of social robots for older adults: A systematic review and meta-analysis. International Journal of Social Robotics, 11, 237–259.

Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. Stanford, CA, USA: CSLI Publications.

Ricoeur, P. (1992). Oneself as another. Chicago, IL, USA: University of Chicago Press.

Scassellati, B., Admoni, H., & Matarić, M. J. (2012). Robots for use in autism research. Annual Review of Biomedical Engineering, 14, 275–294. https://doi.org/10.1146/annurev-bioeng-071811-150036

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–457. https://doi.org/10.1017/S0140525X00005756

Short, T. (2015). Simulation theory: A psychological and philosophical consideration. London, UK: Routledge.

Topol, E. (2019). Deep medicine: How artificial intelligence can make healthcare human again. New York, NY, USA: Basic Books.

Tronto, J. C. (1993). Moral boundaries: A political argument for an ethic of care. New York, NY, USA: Routledge.

Tronto, J. C. (2013). Caring democracy: Markets, equality, and justice. New York, NY, USA: New York University Press.

Turkle, S. (2011). Alone together: Why we expect more from technology and less from each other. New York, NY, USA: Basic Books.

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. Paris, France: UNESCO.

Verbeek, P.-P. (2005). What things do: Philosophical reflections on technology, agency, and design. University Park, PA, USA: Pennsylvania State University Press.

World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. Geneva, Switzerland: World Health Organization.

Zahavi, D. (2014). Self and other: Exploring subjectivity, empathy, and shame. Oxford, UK: Oxford University Press.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. New York, NY, USA: PublicAffairs.