Claudio Alexandre S. Carvalho[*]

Ethical challenges of AI-based psychotherapy. The case of explainability

0. Introduction

Postphenomenology emerged from the examination of how technological artifacts and interfaces mold one's lifeworld, providing a renewed impulse to the transition from the eidetic reduction of phenomena into the acknowledgment of its constitutive relations and mediations. But while important in revealing the impact of technological mediation in the construction of reality, its inquiry provided reduced attention to the complexity of the social framework of design and implementation of technologies, in especially to what is the core of therapeutic mediation, precisely the "in between us" (Van den Eede 2010, p. 157). Over the last decade, the groundbreaking work of authors such as Ihde and Verbeek has been extended to encompass the social and ethical challenges posed by technological artifacts and interfaces molding one's lifeworld, particularly those based on artificial intelligence. In its turn, in its attempt to understand the "material semiotics" (Law 2019) of technological assemblages, Latour's ANT exposed the fluidity of technical events. However, his project failed to recognize how it is precisely the difference between psychic and communicative operations that grounds the emergence stable patterns of interaction. In her recent work, E. Esposito (2022) called for a move from artificial intelligence to artificial communication, showing how much of what we come to qualify as artificial intelligence results from a continuous optimization of information processing oriented to the client's expectations. It is no coincidence that the most recent breakthroughs in AI occurred precisely after attempts to replicate human neurophysiological architecture were replaced by the development of specialized models of information processing.

[*] Post-doctoral research fellow of the FCT (Foundation for Science and Technology) at the Institute of Philosophy of the University of Porto. The research for this article was supported by a postdoctoral fellowship with reference SFRH/BPD/116555/2016.

Scenari, n. 17, 2022 • Mimesis Edizioni, Milano-Udine Web: www.mimesisjournals.com/ojs/index.php/scenari • ISSN (online): 2785-3020 • ISBN: 9791222301167 • DOI: 10.7413/24208914134
© 2022 – The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License (CC-BY-4.0).

We may read this move to artificial communication in light of N. Luhmann's concept of person (1995). Beyond the substantialist models of philosophical tradition, by person the German sociologist understood a form containing expectations and attributions of a certain individual, as the outcome of the co-evolution of psychic and communicative systems. Similarly to Luhmann's model (1995, p. 153) instead of reflexive intelligence, AI interfaces of communication rely on the increasing coordination of artificial agents with the client's expectations as the way to anticipate further sequences of communication. Following Esposito's views, the algorithms of conversational agents are very far from intrinsic awareness of the psychic meaning. As communication technologies they "perform understanding" (2022, p. 10) relying on the development of the ability to identify and mobilize relevant distinctions to a particular addressee. This is possible by virtue of a previous profiling, based on her interactions or context of performance.

***

Although some aspects of general healthcare management may be adapted to psychotherapy, clear-cut criteria guiding biomedical therapies are inadequate to address mental disorders and dysfunctions (Uusitalo, Tuominen & Arstila 2020), as these do not concern natural but "human kinds" (Hacking 1996), embedded in psychological, social, and cultural factors important in their etiology, understanding, and treatment.

There is a considerable particularity to mental health that makes its adaptation of IA models and protocols used in other areas of health services problematic or even unfeasible. These concern very specific ways to identify, classify, and treat problems and dysfunctions. Therapeutic conversation aims for more than an input on one's subjective perception; it is a transformative relational field of moods, feelings, and beliefs. The communicative bond will be decisive in the access and elaboration of problematic nodules, making the client feel heard and understood by another. The reception and containment of unbearable affections favor the reinforcement of that safe space of intimacy and belonging.

AI based psychotherapies may be seen as resulting from a convergence between optimization services increasingly assumed to be requirements for social inclusion (Gori, Le Coz 2007, pp. 73-75) and the emergence of transformative IA, supported by a disseminating network of tracking and monitoring devices integrated with programs for managing performance in intimate and professional life.

Despite being a scientifically informed practice, psychotherapy proceeds through a "management of vague things" (Fuchs 2011), requiring a collaborative "esprit de finesse" argued to be lacking in ear-

lier models and perspectives of AI, supported in the computational model of the mind (Dreyfus 1992, pp. 293-4). However, recent developments in connectionism have defied these restrictions with the formation of neural networks and seed algorithms that are capable of autonomous learning.

Differentiating from a larger AI ecosystem (Winter *et al.* 2021), the integration of AI technologies in psychotherapy -particularly machine learning algorithms and deep learning-, demands a reassessment of the following assumptions: 1) AI simulates human intelligence; 2) it necessarily produces stereotyped diagnoses and treatments; and, finally, 3) by lacking the "common factors" of therapeutic success, it can only assume an ancillary role. These convictions tend to obliterate how AI-specific (or narrow) forms produce transformative effects in the therapeutic medium (Gruetzemacher & Whittlestone 2021).

Along with the detection and prevention of bias resulting from "failed projective identification from humans to machines" (Possati 2021, p. 88; 2022), this requires readdressing both the "translation" of classical therapeutic problems into interactive settings based on AI and emerging ones resulting from alternate modes of diagnosing, categorizing, and predicting dysfunctions.

Some authors noted that along with the traditional bioethical principles (autonomy, non-maleficence, beneficence, and justice), the introduction of IA in clinical practice demands the recognition of a fifth principle: explainability (Floridi 2022, cap. 4).

Explainability has acquired increasing importance to understanding and regulating the use of AI based technology since its innovations are based on complex mechanisms that, although not arbitrary, may be inscrutable. Due to their transformative power, there is a quest for transparency and the provision of an adequate understanding of the criteria and goals central to the ethical assessment of their applications[1]. Explainability is thus a concept that refers to a common social technology, but its interpretation or concretization will differ depending on the observers[2] and context. The various parts interested in understanding the operations of machines require different levels of explanation. End-users of a service and the scientific community must be provided with comprehensive accounts of the functioning and purpose of AI technology, although

[1] Ideally, developers aim for systems that are "complex and intelligent enough to initiate actions on their own, and (…) simple enough to be understandable and controllable by human beings" (Ekbia 2015, p. 63).

[2] "If they lack the relevant technical expertise, a different kind of explanation is needed. This not only reminds us of the problem of education but also leads to the question of *what kind of explanation* is needed and, ultimately, what an explanation is." (Coeckelbergh 2020, p. 121).

a middle level seems required for adequate understanding of its functioning, ensuring conditions for an informed and responsible use. This is all the more decisive when technology is not only disruptive, introducing new ways of accessing and managing information, but also inserted into a field where questions of personal well-being, autonomy, and goals are at the core of interventions. This is clearly the case with psychotherapy.

As a normative principle, explainability may account for the updated prerogatives of free and informed consent, responsibility, and accountability (Coeckelbergh 2020). By virtue of the specificities surrounding the contractual relationship (reduced to perfunctory "terms of service" [ToS]) and therapeutic alliance between machine and client, this line of questioning will reveal the re-emergence of classical problems of psychotherapies, resulting in what Drigalski (1980) characterized as a "system of isolation", reinforcing learned helplessness and dependency (pp. 28ff). These necessitate a reconsideration of the asymmetry between a machine that is supposed to know and the subject, where "algorithmic resonance" provides unwarranted security, which, in some mental conditions, may reinforce the client's withdrawal and isolation. In this paper I will only be able to outline how an adequate form of explainability demands an extension of the concept to encompass various ethical problems related to the "in between" of therapeutic communication.

1. Extension of the concept of explainability

The quest for designing AI agents with operational and functional morality, as conceptualized by Wallach and Allen (2009) is an attempt to complement their increasing assumption of tasks in various fields of activity. Conceived to protect users from harm and malpractice, it goes beyond what J. Moor (2011) called the implicit moral dimension of the machine[3]. Its enforcement may be formal, infusing the agent with clear rules regarding obligations and prohibitions, instilling basic ethical sensitivity, or promoting conditions for autonomous ethical decisions, mainly in the agent's specific contexts of performance. The enforcement of machines' moral behavior by developers, corporations, and government agencies is an attempt to contain liability. However, it is not expected that in the near future, this attribution of moral action and reflection will imply a full sense of moral and legal responsibility.

In these cases, explainability implies that the artificial agent is already

[3] "Computers are implicit ethical agents when the machine's construction addresses safety or critical reliability concerns" (Moor 2011, p. 16).

able to observe certain moral principles and rules, identify and reflect on ethical decisions, and be responsive as to its reasons for taking certain decisions, *i.e.,* be accountable.

Over the last two decades, there has been a growing recognition of the need to assume that AI agents have ethical implications, not only in terms of their ethical impact but also as "explicit ethical agents" (Moor 2011), capable of not only interpreting and following moral guidelines but also providing reasons for their own decisions. This may be interpreted as the logical sequence to Verbeek's call for the recognition of the "moral significance of technologies themselves" (2008, p. 91), extending well beyond the intentions of creators and users. Regarding the algorithmization of ethical reflection, it is important to know if the machine can be the one providing these explanations and, if the answer is affirmative, what are its specific attributions regarding this explainability. In this article, we address a problem that is prior to those questions. We focus on how the explainability of the machine's moral dimension requires a clarification of ethical issues at the heart of a given service or performance, accounting for the generation of expectations regarding the client's experience and action.

Since it concerns the recursivity of communicative forms emerging in a given environment and accounts for the transformative effects that frequently extend well beyond those experienced by the end-users, the ethics of AI cannot be restricted to the devising of conditions to build "moral machines" or agents, even if these will someday achieve a level of ethical excellence in their specific environment.

Ethical and normative discussions on AI technology must attend to their application in different environments. General ethical guidelines for AI, as those adapted from the bioethical discussion in the EU commission's Ethics Guidelines for Trustworthy AI (2019) or in Floridi's recent book (2022: chap. 4) are a symptom of the fading dream of self-regulation. They are having serious difficulties dealing with AI's autonomy, particularly its impact on systems oriented by subjective experience and values.

In her delineation of "Six kinds of explanation in AI", J. Bryson (2019) sustained that, along with an appropriate exposition of "exactly how the system works", which is made comprehensible for the common user of a service, an account of the actions leading to the release of a product would contribute to a better understanding of its purposes, potential, and risks. Such an account concerns all stakeholders, especially those standing at the "sharp end of algorithmic decisions" (Zerilli 2021, p. 177). The move beyond mechanistic explanations was subscribed by L. Possati which underlined the importance of knowing "why and how the AI was created, who were the people who designed it, what their social field and habitus" (2022). Focusing on the case of Replika, he showed the

way conscious and unconscious motivations modulate the performance of an IA conversational agent, unwittingly forming a new kind of artificial unconscious that "inherits" and amplifies some of its creator's anxieties and longings. In that sense, in order to ensure the creation of "friendly AI" we not only have to ensure continuous regulation of utility functions, we also have to account for how those "basic drives" which resist control and correction, are reinforced through their social enactment. As stated by Omohundro (2008, p. 492), "in addition to the design of the intelligent agents themselves, we must also design the social context in which they will function."

Accounting both for the unpredictable outcomes of the introduction of AI agents in certain environments and their learning, the study of "machine behavior" emerged as a new field focusing on the interaction between agents and humans in their respective environments (Rahwan 2019). This transition from the lab into the social environment is inseparable from a closing of the AI Knowledge Gap, supplementing the battery of benchmark tests of a given algorithm with "protocols that access APIs and algorithms 'in the field'" (Epstein *et al.* 2018). Such study requires the inclusion of new disciplines into the AI research community, which may contribute to addressing aspects of the opacity of AI agents that have been largely overlooked. It necessitates broadening the investigation of explainability to include algorithms' interactive potential, especially their enactive and recursive dimensions. This approach is similar to that of critical algorithmic studies, which recognize the algorithms not simply as an operative codification, a way to adequately process information to perform a task, but also account for the way they become enacted by practices in a given environment.

This has been decisive in correcting abstract views of the autonomy of AI systems and agents, where: "environment is understood as a set of features and properties that the agent *senses* and *acts upon*. What is often lost in this conceptualization is the fact that the environment also *supports* the agent in carrying out its actions" (Ekbia 2015, p. 65). An analogy between this way of conceiving IA agents and the view of the "isolated (human) mind" is tempting, since in both cases the abstraction of the agent from the social and normative environment implies the distortion of their performances and potential. When it comes to an AI agent, this necessitates not only taking into account the presumptions underlying its design and implementation but also the way its users come to interact in a given environment.

This is clearly the case in their support of AI agents, because the efficacy of their performance is dependent on the client's or user's observance of contextual and relational norms and expectations, which, while accounted for in their code-source, occur in the "temporal flow of ac-

tion" (Introna 2015, p. 4). Such performances allow for the enactment by the user, which also accounts for previous interactions and their elicited operations. We can conceive of algorithms as non-trivial autopoetic systems, able to learn and refine their processes according to the observations of their outcomes in the environment.

Communicative performances may be observed by agents outside a social system, which should be able to acknowledge their estrangement from the full significance and outcomes of interactions – not only the developers but also researchers, including those from the social sciences and humanities.

Although their main distinctions may be fixed, including the observance of certain moral principles and/or values, therapeutic agents, for instance, are confronted with certain enactions or ways of communicative engagement that, even accounting for the creation of contained contingencies on which most therapeutic interventions rely, may prove unproductive or detrimental. This is especially true when there are no ways to verify access conditions and there are no clear spatial or temporal set limits in the therapeutic setting.

We are extending the notion of algorithm beyond "the boundaries of proprietary software" (Seaver 2017, p. 10) while accounting for their social embeddedness, materialized as tools, interfaces, and environments opened to multiple enactments in social practice. This will certainly imply a correlative extension of the concept of "explainability", addressing dimensions that usually remain outside its scope. An effective explanation of algorithms' workings needs to address their larger impact on the environment and, taking into consideration that they may, autonomously or under monitoring, originate typical problems, it has to entail ways in which these may be prevented, resolved, or amended. It may be understood as a reaction to the various declinations of ethical washing (Yeung *et al.* 2020) as strategies to evade effective regulation and future legal responsibility. That is particularly relevant in the implementation of mental healthcare protocols, where the transition from technical and controlled laboratorial studies of AI agents and their insertion in end-user interfaces obliterates rigorous certification (Hiland 2021).

The call for reform is bolstered here by the predominance of the concept of transparency in psychotherapy, which includes an epistemic and normative dimension that are critical in its marketing, scientific status, and regulation. It is an attempt to explore a larger framework of a concept that, if taken seriously, provides an overview of the design, development, and implementations of a given social technology. It has different modes and goals depending on whether it is part of the self-observation or inner assessment of a given service provider or is directed to its stakeholders.

Resorting to therapeutic semantics, Coeckelberg (2019) argued for a relational justification of explainability that decenters the observation of AI's technologies from experts. AI applications must be assumed to be moral agents in relation to the moral *patiens* affected by them. The technical dimension of explainability, particularly its epistemic condition, has to be articulated through the ethical obligations of accountability and answerability. Besides the basic warranties of safety and control, in assessing IA technologies, there is "the obligation to greater awareness of unintended consequences and the moral significance of what they do" (2019, p. 16). This is critical for therapeutic interfaces because it calls for particular sensitivity to the potential uses and misuses of a technology according to its addressees or users, based on their typical capacities and situations.

Explainability depends on the promotion of digital literacy on the part of end users. In some cases, this task has been assumed by governments and non-profit organizations, which provide information on websites, at public events, and even in courses aimed at various levels of expertise and use of AI[4].

The problem of control begins with undetectable or subliminal unconscious projections on the machine, affecting the efficacy of instructions and experiences. Simultaneously, it raises the issue of encouraging value reinforcement in new interfaces (Verbeek 2008). Social biases are entrenched in the larger social environment and may be easily reproduced and amplified by those interfaces. They cannot be reduced to glitches or flaws. They are signs of the inexorable functioning of AI machines. In some cases, their functioning makes bias more evident, creating new communicative conditions for their awareness and change.

Regarding explainability, it has been argued that it may be incorrect to assume that IA-based decision processes are more opaque than human decision-making processes. A more transparent exposition of the grounds for AI decision-making may itself contribute to better understanding the actions in "traditional" systems (Zerilli *et al.* 2021, p. 41). Similarly to B. Kuipers's (2012) view of collective institutions and social systems as an evolving *genus* of artificial agents provided with sensory, representational, and deliberative operations oriented towards specific outcomes, we may consider algorithms as the ultimate refinement and condensing of a computational function that was prepared by the "natural" evolution of social systems.

[4] One of such examples is elementsofai.com provided by the University of Helsinki.

2. The potential of therapeutic chatbots

A chatbot is generally defined as a computer program that engages in natural language conversations with other agents. It is able to recognize verbal expressions and certain pictographic signs and expresses itself through verbal messages, whether written or spoken. While its distinction to the larger genus of virtual embodied conversational agents is becoming increasingly blurred, the chatbot's limited visual and physical presence may be considered its *differentia specifica*.

Therapeutic conversational agents differ from chatbots that perform routine tasks such as customer service and "personal agents" that aim to provide general assistance (Alexa, Siri, or Cortana) oriented by objectivity and clarity of communication. Resorting to a common maritime metaphor, their companies and sponsors present them as assisting users in navigating mental health issues and well-being challenges.

The impact of therapeutic conversational agents is far from being restricted to the way the user relates to herself, not only in the sense that, as a technology of self-thematization, it alters the way the subject perceives and acts on her environment but also because it generates new expectations in various social systems. Some authors refrain from considering the chatbots as agents; however, the singularity of their communicative action is decisive for the generation of a therapeutic bond, extending well beyond the identification with "one's" algorithm in other types of service (Colbjørnsen 2016).

In creating ELIZA, the first ancestor of today's chatbots, Joseph Weizenbaum (1966) parodied the non-interventionist therapies of Carl Rogers' personalism. But he found that his script, which conjured up the expression of one's burdens and anxieties and established ways to acknowledge and approve them, led many users to form an intense emotional bond with the machine. A liberating experience was reported even by those who had followed its development, knowing not only that the lines came from a computational process but also that it was based on the predetermined management of some recurrent topics of therapeutic conversation and the mirroring of the client's expressions, designed to provide credible approval of one's disclosures and instigate further confidences. One of the highlights of what would become known as the "ELIZA effect" was when, a few minutes after starting to interact with the chatbot, Weizenbaum's own secretary asked him to leave the room (Weizenbaum 1976, pp. 3-4). Later, Weizenbaum would confess his perplexity: "[w]hat I had not realized is that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people" (Weizenbaum, 1976, p. 7). In order to prolong their interaction with the program, users were frequently willing

to overlook its errors and limitations. Considering the efforts to develop a computer program able to formalize a psychiatrist's intervention, Weizenbaum worried that such an endeavor would imply tailoring the therapeutic setting to the confines of computational processing and thereby disregarding the client's needs (Weizenbaum 1976, p. 6). But at a deeper level, he feared a moral trivialization of the human being, which was based on a lie, summarized in the computer's use of the personal noun "I" (Turkle 1995, p. 106). In a chapter entitled "Against the Imperialism of Instrumental Reason," he reaffirmed his rejection of a computer-based form of therapy "not on the grounds that such a project might be technically infeasible, but on the grounds that it is immoral" (p. 266).

Kenneth Colby[5], a psychiatrist at Stanford University, took the possibility of computerizing psychiatry seriously, making it his lifelong scientific pursuit. He created SHRINK with the "intent to help, as a psychotherapist does, and to respond as he does by questioning, clarifying, focusing, rephrasing, and occasionally interpreting." (Colby *et al.*, 1966, p. 149). In fact, he famously sustained that "A human therapist can be viewed as an information processor and decision maker with a set of decision rules which are closely linked to short-range and long-range goals" (p. 150)[6].

ELIZA is a prime example of a handcrafted program, largely consisting of a parsing of written input that, by establishing connections with predetermined keywords, provided corresponding outputs. Additionally, in its formation of expressions, it relied on anaphoric references already inserted by the user, particularly pronouns, providing a sense of the interlocutor's engagement. In cases where the correspondence with a keyword in the database couldn't be established, in a very Rogerian way, it responded with new queries or questions, what is today known as the "ultimate default category": "please tell me more…", "very interesting. Please go on", "can you thing of a special example?"[7]. But it lacked any way to process the context of the conversation or a structured memory of the user's states.

The critical observation of the ELIZA effect has pointed to the user's willingness to fill in the blanks of the program, obliterating how the limitations of the machine may promote the development of one's articulation of self-understanding and new ways for its expression. The perceived

[5] Also creator of PARRY, a program simulating a schizophrenic patient, which had famous exchanges with ELIZA. E.g.: https://datatracker.ietf.org/doc/html/rfc439.

[6] "Almost everyone who has participated in these dialogues reports that he comes to feel annoyed and frustrated by the program's responses" (Ibid).

[7] Here it was also possible to introduce backchannels such as "yeah", "sure", "right", or expressions such as "uh-huh" which could be considered verbal analogs to nonverbal cues such as nodding.

failure to acknowledge the user's accounts may spark the search for more effective ways of expressing oneself, renewing the self-reflective exercise, and seeking recognized action.

"Enaction" refers to the opportunities generated by a specific exchange infused with expectations that must prove efficacious in subsequent interaction in therapeutic settings based on written communication. At the most basic level, it is dependent on the sensory perception of material signs, followed by the interpretation of their symbolic meaning. As it occurs in general forms of psychotherapy, communication forms emerge based on the constitutive distance between "sender" and "receiver" or "alter" and "ego," the in between. In the elaboration of a meaningful message to the other, the subject has to assume herself as the object of interpretation, a reflexive observation that may or may not be accepted.

When the narration of oneself is conceded space and met with the manifestation of genuine interest, the other is confirmed in his role as assistant or facilitator of a process of self-discovery. This communicative framework is enactive in the sense that some opportunities regarding the exercise of self-thematization, accompanied by emotional, cognitive, and behavioral levels, become available with the development of new awareness of one's prospects. In therapeutic communication, the previous themes of conversation may be recovered and revised in light of new interpretations and insights.

Communicative forms of therapy promote second-order observation, i.e. the observation of one's observations and their operating distinctions. This enables their reframing of a current situation or problem, as well as their emotional correlata. Although it emerges from the reaction against repressive pressures of conformity, contemporary therapy may be seen in the context of modernity's "increasing resort to rituals of self-knowing" (Hahn 1982, p. 407). Unlike confession, its work does not necessarily aim for what the subject knows following the interiorization of interdictions, but something she has not yet realized about herself, or, in terms common to psychoanalysis and sociological inquiry, something she does not know that she knows.

The therapeutic relationship's asymmetry is reframed in the new conditions of AI interfaces, but it is still based on the idea of a pact evolving co-responsibility in communication, a responsive practice.

Various therapeutic currents rely on a kind of biographizing of the individual, constructed in view of the present account of a difficulty. Interaction in conversation invites the elaboration of how certain elements affect the individual. It may provide the relief observed when addressing painful subjects or memories, but also the feeling of being understood and accompanied by someone who cares. In cases of anxiety and depression, this acknowledgement of the impact of episodes that menaced or

disturbed one's sense of integrity allows space for a transformative kind of remembering. Instead of reenacting the traumatic episode deemed to be at the origin of one's suffering, the joint exploration aims for a reintegration of these memories into a new coherent narrative of oneself. This is clear in interfaces relying on conversational agents. Their acknowledgement and understanding, manifesting interest in their articulation, allow indications and clarifications to be more or less suggestive, depending on the therapeutic setting. But the anticipation of latent psychic processes evolving through these joint elaborations, similar to what occurs in traditional psychoanalysis, is outside the scope of a chatbot. Except for standardized services where behavioral and contextual profiling are highly optimized[8], the anticipation of emerging (non-explicit) quests or ideas would require the awareness of non-verbal language and the reliance on metaphoric construction of meaning.

According to B. Christian (2020), the problems with AI applications stem from their inexorable exactness, which poses multiple challenges in making certain goals explicit, detecting and correcting bias, and acknowledging certain norms and values. This poses the tentative possibility of programming AI to learn from its own operations or experience in order to acquire greater sensitivity to problems whose complexity resists codification. The model of learning through simple imitation and replication of behavior was too limiting, making interfaces and agents to deal with complexity and contingency. At the same time, an exhaustive codification of conditions of performance is impossible.

Reinforcement learning [RL], particularly in genetic algorithms and neural networks, emerged as a way to achieve a better mapping of a program's environment and recursively assess the results of their interventions. Instead of proceeding from pre-established procedures or decisions according to stable decision trees, machine learning implies autonomously learning from data sets, generally after multiple training cycles. However, in IA therapeutic chatbots, it must begin with the method of data collection and organization. Learning from experience enabled new, dynamic ways to simulate how to achieve desired outcomes without settling for partial achievements. The retrieval of probabilities allowed the actualization of representations for action according to the establishment of decision rules. In order to attain such an operation, we have to consider the specificities of the therapeutic environment and the pitfalls of its communication. These relate to the ways of engaging in an activity whose rewards, aside from their sparsity, are difficult to identify and achieve. A qualified form of interest involvement seems to be es-

[8] Regarding how ordinary virtual assistants anticipate clients' explicit requests, see: Youyou, Kosinski, Stillwell 2015.

sential for therapeutic intervention, as it motivates one to partake in the patient's experiences and decisions. However, in order to avoid devolving into mere curiosity, such as by encouraging erratic forms of dramatization, it must be guided by therapeutic values and goals.

Until the turn of the century, chatbots performed mostly based on an exhaustive codification of all the anticipated inputs and the transformative rules to apply into parsed strings of signs. Due to the conflicting application of rules and conditioning, as well as the difficulty of updating its operative coding, it would be difficult to provide adequate outputs to new, unpredicted inputs and sequences. A timely updated inference of the context and adequate output would require an immense computational load to establish what Minsky termed the "frame of action": dialogue history, task record, encoding of the domain, norms of interaction, and updated information on the user's intentions and goals. The inclusion of statistical data-driven systems, particularly RL, enabled the optimization of recognition of natural languages, multi-level management of dialogue, and natural language generation. Deep neural networks have recently emerged, allowing some tasks of output responses to be generated without modular processing of the input, known as "Seq2Seq" (Sutskever, Vinyals, Le, 2014). The systems are able to fulfill immediate tasks while maintaining orienting goals and higher conditioning of sequences, such as strategic management, for instance, by adopting an engaging or defiant tone according to the present state.

The various forms of observation mobilized in therapy take the illness or dysfunction as the marked space that grounds prospective actions and future assessments of the patient's state. According to their specific distinctions, "therapeutic conversation is about developing a shared idea about the forms of the patient's distinctions and indications, addressing his role in the creation and maintenance of a symptom" (Simon 2015, p. 288). Fictions of understanding play a crucial role in enabling the transposition of opacity into the transparency of the treatment of symptoms or burdens, which may therefore be considered and re-authored.

New chatbots have a greater capacity to retain and organize information to be retrieved according to the dialogue state tracking, replacing an exhaustive mapping of the conversational history and context. This openness to the user's contingency, as generated throughout the interaction, favors the exercise of self-thematization according to the present sequence of acquisitions. Instead of an exhaustive processing of information regarding the individual (and interactions), algorithms select key themes and problems that may occasion new interaction sequences. In that sense, their generativity is anticipatory, proceeding in view of the opportunity to introduce new schemes of self-understanding and observation. Here, the greatest progress in the new generation of chatbots

consists in their maintaining different states (with corresponding output sequences) and updating their probabilities according to new inputs and interactions. In that sense, it largely exceeds the shallow use of anaphoric reference, which we found in earlier computer programs such as ELIZA or SHRINK. The acceptance of input by automatic speech recognition (NLP) leads to dialogue management, which relies on categories to be addressed and scripts that prompt new sequences or output, generally controlling the dialogue flow (Traum 2017). Nonetheless, their effective treatment of the contingency formed in interaction means that their outcomes cannot be previously available to any of the participants in communication (Esposito 2022, pp. 9-10).

Among the new computational operations introduced by the statistical models is the management of various beliefs concerning the current state of the dialogue, particularly regarding the "real" meaning of the user's statements. In this sense, the concept of states being "partially observable" and proceeding according to probability distribution is fundamental to dialogue management[9]. Even under ideal conditions of complete disclosure, the system cannot be certain that it has correctly identified the user's states or intentions. Instead of matching a pre-established meaning, probabilistic grammars and parsers proceed to a parallel formulation of possibilities. Therefore, previous communicative sequences are organized in such a way that they remain open to subsequent interpretation and may be updated in new interactions (they will be evaluated according to Q-function). This means that this communicative mediation works on the latency of present states, the development, and the selectivity of the person of reference, *ie* the client, allowing for flexible points of entry where the burdens may be observed from new perspectives.

A frequent assumption when considering AI-powered mental health agents such as humanoid robots, embodied virtual agents, and chatbots is that they are necessarily conceived to replicate or simulate intelligent human behavior. This is undeniably true because intervention necessitates proficiently using natural language, including stereotyped ways of treating others and forming an empathic bond with the user. However, considering their specific "nature" and potential, including different modes of gathering, organizing, and retrieving information that may provide a more detailed account of the client's individuality, that perspective seems to call for some qualifications. That more detailed analysis of the individual and the more thorough consideration of its possibilities may lead to the execution of strategic interventions whose causal grounds

[9] Particularly in the Partially observable Markov decision process (POMDP) combining regular Markov Decision Process to model system dynamics with a hidden Markov model that connects unobservable system states probabilistically to observations.

may be difficult to explain to a human observer. This does not necessarily imply structural changes in current therapy methods, though these may be envisioned, opening the door to a "regression" into oracular or charismatic modes of communication (Macho 1999).

Through refined analysis of an increasing number of cases, AI has the potential to create new categories to classify and understand suffering, as well as more efficient ways to intervene. However, because these new classifications concern the human experience of suffering, which can be treated in various ways, the assessment criteria and treatment options must be explainable in terms appropriate for different audiences. This may be the case, for instance, when the "costs of change" or "correction" (Luhmann 2019, pp. 39-40) of a problem are disproportional to the projected benefits or lead to worse problems, so that the efforts one must undergo to achieve a resolution may be counter-productive. Evaluation of online behavior patterns also allows for detecting upcoming disturbances or predicting mental problems. But taking into account that these are "just" predictions, should the program, by default, advise the client or related persons?

Explainability must answer the concerns of at least the users, the government and regulatory agencies, and the scientific community that can assess the technical aspects of a particular program. These are the main vectors to consider when implementing effective answerability and accountability. However, for explainability to be effective, it must consider the impact of a given program interface in the larger environment and the enactments it may occasion. We have good reasons to believe that narrow forms of AI, *i.e.* those whose computational abilities and design are oriented to a particular task or challenge, may enable greater control and epistemic ways to assess the ethical impact of a technology. Mental health settings based on AI pose a particular challenge since, while they answer the problems and aims of individuals and groups through protocols that have been standardized, to communicate in this system, a therapeutic agent is required to have a common knowledge of a wide domain of subjects. At the same time, "therapeutic culture" seems to be potentially applicable in every social system.

In fact, we can't help but notice a significant restriction and change in the factors deemed common in psychotherapy, all of which, following Wampold's mapping (2015), concern the "in between" therapist and patient. Explicit convergence may be achieved in developing the therapeutic alliance, the goal consensus (and positive regard) of treatment, and its cultural adequacy. Nevertheless, the therapeutic work and support provided by an IA agent lack the depth and identification required for empathy and genuine mentalization, even if the client may feel these. The lack of emotional resonance and embodied empathy alters the quality of relationships and therapeutic outcomes.

However, we should not dismiss a therapeutic agents such as Woebot or Wysa simply because it keeps the door to the other scene of the unconscious closed. Given its limitations in terms of range and its promptness in providing solutions that are supposed to reframe complex situations, this may indeed be the wisest choice.

On the one hand, the chatbot's absence of emotional resonance and density ensures a non-judgmental position, promoting revelation and attenuating the management of impressions on the other (Ho *et al.* 2018). At the same time, this absence of resonance is associated with a sense of superficiality or instrumental use that differs from authentic listening and recognition, which aims understanding and identification with the other's situation. Instead of considering the client's hypothetical mental states, the virtual agent concentrates on the action potential. This is the fundamental limitation of these agents. They process multiple possibilities and select according to detected patterns of dysfunction, and the devising of appropriate expressions in the form of questions, indications, and suggestions. Their responsive practice lacks identification with the other's experience, a pulsating form of empathy and interest that allows depth and sensitivity to thought processes.

As a business model, AI-based interfaces for psychotherapy are engaged in a systematic attempt to suppress any intermediary between the user and the program. This suppression of the mediation of any mental health expert at every step of the therapeutic process could only be accomplished through a tactful relationship with governments and health authorities. Instead of a tool or program to be administered under the guidance of a clinician, which would necessarily undergo all the tests on their safety and effectiveness for the users and the larger public, AI therapeutic interfaces fly under the radar, categorized by regulatory health authorities as harmless "mobile applications" (Hiland 2021, p. 22). In order to circumvent the impositions that regulatory agencies such as the FDA impose on health-care providers, including training and deontological obligations, chatbot enterprises maintain that they do not provide diagnostics or treatment but rather assistance and support. This is clearly refuted by the practice of most therapeutic chatbots, which provide an assessment of the user's mental health in view of relief or improvement through exercises, but also by their own marketing of the apps (not only in ads but also in scientific articles asserting its efficacy).

All responsibility is transferred to the user, which necessitates a robust concept of explainability that extends beyond the routine signing of ToS and informs on the multiple risks of these interfaces and the current dispositions to minimize them.

Chatbots' hybrid nature is expressed in their self-presentation as not therapists or psychologists but companions, buddies, or allies in men-

tal health management, serving people and businesses. The question is whether, by denying Woebot therapist status, the enterprise limits not only the scope of its action and impact on the mood of the customer or patient but also its liability for negative consequences associated with its use.

Explainability plays a decisive role in the social promotion of a given technology, positioning it in the commercial and regulatory markets and providing evidence (in various forms) of its security, fairness, accuracy, and efficiency (Babuta *et al.* 2018, p. 18). In AI-based therapy, explainability may not be seen as a deterrent to innovation but as a decisive aspect to reinforce therapeutic alliances and building trust in its setting. At the same time, going beyond its technical specifications and recognizing its transformative effects on the environment may help prevent risks related to moral dysfunctions and their inadequate use by moral patients. Such extension demands reflexive practice, aware of the various pressures that may affect therapeutic service.

Concluding remarks

At this point in the evolution of therapeutic techniques, the human's exclusive contribution is the assurance of a lived resonance of one's words and signs. Rather than a calculation of the spontaneity of empathy, trained therapists achieve a density of interpretation of the patient's affects and experiences, a communicative attunement that allows for a better direction of interventions. This is expressed in the sensitivity to non-explicit processes that must be welcomed and interpreted, requiring openness to the "analogic" dimension of human behavior, recognizing non-verbal cues, speech rhythm, and tone, strata whose full scope remains intangible to AI agents.

However, AI agents have specific advantages associated with their computational abilities. These concern communicative abilities that are severely limited in human forms of cognition, as in the case of predictive accuracy based on the detection of patterns or the ability to better retrieve the evolution of key themes appearing in dialogue. At the same time, the conversation with an AI agent favors the client's disclosure (absence of a moral judgment) and fewer worries about monitoring the presentation of the self. We must also acknowledge the greater scalability and affordability of AI protocols and their ability to reach people with limited access to therapeutic care.

It was the recognition of AI therapeutic agents' potential to collect, organize, and interpret data that led designers and developers to restrict their aims for higher achievements. Along with the increasing potential in the recognition of emotions, the observation of an intense

bond of trust and reliance on support implies the acknowledgement of the powerful and possibly detrimental influence exerted by the agent's expression and suggestions. This development of a narrow form of artificial intelligence may be understood as part of a larger movement toward AI's domestication (Kempt 2020). What was lost in creativity was gained in security and, according to the critics, in programs that conformed the desires to the expectations of the current capitalist society. We may not fully exclude the possibility that some limitations may be inherent to the therapeutic current adopted by a specific program, for instance, CBT, which refrains from deepening introspection or the interpretation of the past.

It is through the new relational forms of meaning construction that renewed doubts arise as to the methods and goals of therapy. These concerns, of a technical and moral nature, may benefit from the extension of the concept of explainability that we have explored in the present article. The elaboration of explanatory models of specific treatments in mental health was already at the heart of epistemic validation of psychotherapies, with increasing calls to use them to regulate and oversee their protocols and practice. Valuing the long observation of the recurrent ethical problems of traditional therapy, we may reinforce safer and more reliable forms of AI based therapies. By privileging the user's understanding, this approach presents a valid alternative both to the ethical washing of AI enterprises and corporations and the paternalistic approaches based on strict governmental regulation.

In implementing AI powered therapeutic settings, these specificities play a decisive role in assessing various ethical questions at the core of mental health services. In addition to clarifying its priorities in light of their technical constraints, a complete model of explainability must take into account these factors as well as how they affect their "moral patients" environments. This implies a sensitivity to their demographic target and typical difficulties, frailties, and impairments, which make some misconceptions and misuses of the setting more likely.

Bibliography

Babuta, A., Oswald, M., Rinik, C.
2018 "Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges," Whitehall Reports, Royal United Services Institute, London.

Christian, B.
2020 *The Alignment Problem*, Norton & Company, New York.

Coeckelberg, M.
2019 "Artificial intelligence, responsibility attribution, and a relational justification of explainability". *Sci Eng Ethics* 26:2051–2068.
2020 *AI Ethics*, MIT Press, Cambridge, Mass. and London.

Colby, K., James, M., Watt, B., Gilbert, J. B.
1966 "A Computer Method for Psychotherapy: Preliminary Communication", *Journal of Nervous and Mental Diseases* 142, 2: 148-152.

Colbjørnsen, T.
2016 "My Algorithm: User Perceptions of Algorithmic Recommendations in Cultural Contexts". Selected Papers of AoIR2016: The 17th Annual Conference of the Association of Internet Researchers.

Dreyfus, H. L.
1992 *What Computers Still Can't Do. A Critique of Artificial Reason*, MIT Press, Cambridge.

Drigalski, D. v.
1980 *Blumen auf Granit: Eine Irr- und Lehrfahrt durch die deutsche Psychoanalyse*. Ullstein, Frankfurt/Main.

Ekbia, H. R.
2015 "Heteronomous Humans and Autonomous Agents: Toward Artificial Relational Intelligence", in: eds. J. Romportl, E. Zackova, J. Kelemen, *Beyond Artificial Intelligence. The Disappearing Human-Machine Divide*, Springer, Cham.

Epstein, Z. et al.
2018 "Closing the AI Knowledge Gap", *arXiv*, 1803.07233.

Esposito, E.
2022 *Artificial Communication. How Algorithms Produce Social Intelligence*. London/Cambridge, Mass: The MIT Press.

European Commission.
2019 Ethics Guidelines for Trustworthy AI.
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Floridi, L.
2022 *The Ethics of Artificial Intelligence. Principles, Challenges, and Opportunities*, Oxford University Press, Oxford.

Fuchs, P.
2011 *Die Verwaltung der vagen Dinge, Gespräche zur Zukunft der Psychotherapie*, Auer, Heidelberg.

Gori, R., Le Coz, P.
2007 "Le coaching: main basse sur le marché de la soufrance psychique", *Cliniques méditerranéennes*, 75: 73-89.

Gruetzemacher, R., Whittlestone, J.
2021 "The transformative potential of artificial intelligence". (arXiv:1912.00747v3).

Hacking, I.
1996 "The looping effects of human kinds", in: Sperber D, Premack D, Premack AJ, eds. *Causal Cognition*, 351-383. Oxford University Press, Oxford.

Hahn, A.
1982 "Zur Soziologie der Beichte und anderer Formen institutionalisierter Bekenntnisse: Selbstthematisierung und Zivilisationsprozeß", *Kölner Zeitschrift* für *Soziologie und Sozialpsychologie* 34, 3: 407-434.

Hiland, E. B.
2021 *Therapy Tech. The Digital Transformation of Mental Healthcare*, University of Minnesota Press, Minneapolis.

Ho, A., Hancock, J., Miner, A. S.
2018 "Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot", *J. Commun*. 68(4): 712–733.

Introna, L. D.
2015 "Algorithms, governance, and governmentality: on governing academic writing", *Sci Technol Human Values* 41(1): 17–49.

Kuipers, B.
2012 "An Existing, Ecologically-Successful Genus of Collectively Intelligent Artificial Creatures". *Proceedings ICCCI*. https://doi.org/10.48550/arXiv.1204.4116

Law, J.
2019 "Material Semiotics"
http://www.heterogeneities.net/publications/Law2019MaterialSemiotics.pdf

Luhmann, N.
1995 "Die Form 'Person'", in: ed., Id., *Soziologische Aufklärung 6. Die Soziologie und der Mensch*, 142-154, Westdeutscher Verlag, Opladen.
2019 [1975] "Strukturauflösung durch Interaktion. Ein analytischer Bezugsrahmen", in: Id., eds. Lukas, E., Tacke, V., *Schriften zur Organisation*, vol. 2, 29-58, Springer, Wiesbaden.

Macho, T.
1999 "Zur Ideengeschichte der Beratung. Eine Einführung", in: ed. Prechtel, G., *Das Buch von Rat und Tat*, Diedrichs, München.

Moor, J. H.
2011 "The nature, importance, and difficulty of machine ethics", in: Anderson, M., Anderson S. L., *Machine ethics*, 13–20, Cambridge University Press, Cambridge.

Omohundro, S.
2008 "The Basic AI Drives", in: eds. Wang, P., Goertzel, B., Franklin, S., *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483–92, IOS Press, Amsterdam.

Possati, L. M.
2021 The Algorithmic Unconscious How Psychoanalysis Helps in Understanding AI, Routledge, Oxon.
2022 "Psychoanalyzing artificial intelligence: the case of Replika". *AI & Soc.*

Rahwan, I., Cebrian, M., Obradovich, N. et al.
2019 "Machine behavior", *Nature* 568: 477–486.

Seaver, N.
2017 "Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems", *Big Data Soc* 4: 1–12.

Simon, F. B.
2016 [1993] "Die andere Seite der Krankheit", in: ed. Baecker, D.: *Theorie der Form*. Suhrkamp, Frankfurt/Main.

Sutskever, I., Vinyals, O., Le, Q.
2014 "Sequence to sequence learning with neural networks", *Advances in Neural Information Processing Systems*, 3104-3112.

Traum, D.
2017 "Computational approaches to dialogue", in ed. Weigand, E., *The Routledge Handbook of Language and Dialogue*, 143–161, Taylor & Francis, New York.

Turkle, S.
1995 *Life on the Screen: Identity in the Age of the Internet*, Simon & Schuster, New York.

Uusitalo, S., Tuominen, J., Arstila, V.
2020 "Mapping out the philosophical questions of AI and clinical practice in diagnosing and treating mental disorders". *J Eval Clin Pract*. 1–7.

Van Den Eede, Y.
2010 "In between Us: On the Transparency and Opacity of Technological Mediation," *Foundations of Science*, 16, 139–159.

Verbeek, P.-P.
2008 "Morality in Design. Design Ethics and the Morality of Technological Artifacts", in eds. Vermaas, P.E. Light, A., Moore, S. A., *Philosophy and design: From engineering to architecture*, 91–103, Springer, Dordrecht.

Wallach, W., Allen, C.
2009 *Moral machines: Teaching robots right from wrong*, Oxford University Press, New York.

Wampold B. E.
2015 "How important are the common factors in psychotherapy? An update", *World Psychiatry* 14: 270–77.

Weizenbaum, J.
1966 "Eliza – a computer program for the study of natural language communication between man and machine", *Communications of the ACM,* 9(1):36-45.
1976. *Computer Power and Human Reason*, W. H. Freeman, New York.

Winter, N.R., Cearns, M., Clark, S. R. *et al.*
2021 "From multivariate methods to an AI ecosystem". *Mol. Psychiatry*, 1-5.

Yeung, K., Howes, A., Pograbna, G.
2020 "AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing", in eds. Dubber, M., Parsquale, F., Das, S. *The Oxford Handbook of Ethics in AI*, 77-106, Oxford University Press, Oxford.

Youyou, W., Kosinski M., Stillwell, D.
2015 "Computer-Based Personality Judgments Are More Accurate Than Those Made by Humans." *Proceedings of the National Academy of Sciences* 112 (4): 1036–1040.

Zerilli, J.
2021 *A Citizen's Guide to Artificial Intelligence*, MIT Press, Cambridge, Mass.

Ethical challenges of AI-based psychotherapy. The case of explainability

This paper examines the reemergence of some of the traditional ethical issues of psychotherapy in therapeutic interfaces resorting to AI-driven conversational agents. I will begin by 1) proposing the extension of the operatory concept of explainability to encompass ethical problems related to the "in between" of therapeutic communication. Then, 2) I attend to the evolution of therapeutic chatbots, and how, as self-thematization media, they optimize an algorithmic resonance with the problems of their person of reference. For this, I argue, instead of resisting the contingency of the users' inputs, chatbots rely on it to create new generative distinctions. I conclude that a consistent explainability of AI-driven chatbots needs to move beyond the clarification of algorithmic mechanisms to address the potential effects of this technology of self-thematization.

KEYWORDS: AI Ethics; Therapeutic communication; Self-thematization; Transformative effects; Explainability.