On the algorithmic unconscious:

Can we humanize AI with psychoanalytic principles?

Duc-Hung Nguyen[1], Manh-Tung Ho [1, 2*]

1. Institute of Philosophy, Vietnam Academy of Social Sciences, 1 Lieu Giai Str., Ba Dinh Dist., Hanoi, 100000, Vietnam
2. Centre for Interdisciplinary Social Research, Phenikaa University, Yen Nghia, Hanoi, 100803, Vietnam

Email: tungmanhho@gmail.com

<< Draft Manuscript VNAIethics-20250207 >>

> Fully content, Crow 2.0 thanks Kingfisher and eagerly returns home. The Crows, however, find this Crow 2.0 to be a scammer, for he looks nothing like Crow but still claims to be in the family. They all gather to chase him away. It takes several beatings for the crow family to believe this is indeed a family relative. Crow 2.0 is shocked…
>
> - In Contentment, *Wild Wise Weird: The Kingfisher Story Collection, Vuong (2025)*-

Figure 1: Artwork created by DALLE: A surreal depiction of AI and the unconscious mind. A majestic kingfisher with a fragmented, dreamlike form—half mechanical, half organic.

Humanizing AI is one of the most pressing issues in the development and use of artificial intelligence in recent years, and many scholars have highlighted the critical need for emphasizes humanistic values as a core foundation for developing and using AI for societal goods. Here, human subjectivity should be in the forefront AI development and integration, because "while algorithmic knowledge of humans can be vast and can outperform their own knowledge, it remains foreign to their subjectivity", noted Razinsky (2023, p.394).

Numerous research works have attempted to apply the achievements of social sciences and humanities in making sense of AI development. Among these is *Humanizing Artificial Intelligence: Psychoanalysis and the Problem of Control*, edited by Fabio Tollon Possati, who pioneers applying psychoanalytic theories on AI and algorithms. Possati refers to this as *"an odd couple"* (Possati, 2021), a pairing he has explored in depth because he believes that focusing on neuropsychoanalysis and affective neuroscience, rather than solely on cognitive science, could yield more positive outcomes in AI development (Possati, 2021).

Overall, this work examines the intersection of artificial intelligence (AI) and psychoanalysis, considering the subject matter and research methods of both fields. The investigation contributes various innovative perspectives on humanizing AI while also addressing the challenge of evaluating and managing technological development. Therefore, it can be said that this work is not only about AI or defending the assumptions of psychoanalysts but represents a synthesis of philosophical, psychological, ethical, and technological issues surrounding AI development in society.

The five chapters of the work revolve around a single primary goal: proposing a development path for the *relationship* between humans and AI around the principles of psychoanalysis. The contents offer a right blend between how AI must evolve to reflect humanistic qualities in its relationship with humans (chapters 1, 3, and 5) and how humans have changed and need to continue changing to adapt to today's technological world (chapters 2, 3, and 4). Issues central to human subjectivity such as pleasure, libido, imaginary, perceived fairness, and the striving toward true objective reality with new technologies are extensively discussed.

In Chapter 1, *A Freudian implementable model of the human subject*, Paul Jorion proposes a new approach for advancing AI toward artificial general intelligence (AGI) without the need to create artificial consciousness (AC), drawing on Freud’s metapsychology. Paul provides an overview of several key concepts in Freud's psychoanalysis, such as Libido, the Pleasure Principle, the Mechanism of Repression, and the three-part structure of the psyche (ego, super-ego, and id). Based on this, the author presents a rather controversial conclusion regarding the solution to saving humankind. He argues that Plan C (humans being replaced by robots) is more feasible than Plan A and Plan B (saving humankind or settling on other planets). Specifically, instead of creating "intelligent robots," we should focus on establishing "human subjects" based on Sigmund Freud's psychoanalytic theories—“a masterpiece of scientific achievement” (p. 28).

Chapter 2, *Psychoanalysis and artificial intelligence: Discontent, disruptive algorithms, and desire*, written by Hub Zwart, discusses two questions, which are “How can psychoanalysis contribute to coming to terms with AI and to what extent does AI allow us to update psychoanalytic theories of the unconscious?" (p. 29). This discussion is based on the Freud's idea of the psyche and the Lacan's endeavor to expand this idea. Resolving the two questions above leads the author to two conclusions. First, technology in general, and AI in particular, is created as a filter to control the large amounts of information from "the raw real" that affect human physical and psychological systems. However, these technologies themselves are becoming an independent source of information impacting humans (information overload). The second conclusion is the antithesis of Yuval Harari's conclusions in "Homo Deus". Both Lacanian psychoanalysis and Harari's perspective point out two ways for humans to confront the real: "the Imaginary" (stories or mythical figures) and "the Symbolic" (algorithmic). While Harari argues that "the Imaginary" is redundant for addressing modern-day issues and that humans should focus only on "the Symbolic," Lacan's perspective demonstrates the necessity of "the Imaginary" through the concept of "sublimation", reconciling the symbolic and the imaginary, through products such as poetry, architecture, and so on.

In the next chapter, *Nothing human is alien - ai companionship and loneliness*, Kerrin A. Jacobs analyzes the use of AI to address human loneliness in today's increasingly digitized society. The most notable aspect of this chapter is the author's use of Freud's psychoanalytic theory to explain the origin of human loneliness from a socio-pathological perspective. Both individual pathology and cultural pathology are explained based on the theory of repression mechanisms, meaning the use of libido energy under the imposition of culture, and sublimation, meaning the use of libido energy in the process of cultural production. Based on this theory, each individual feels lonely when cultural norms influence them to the point where they are overly repressed, expending too much energy in cultural production and attempting to meet cultural demands. Therefore, according to the author, an individual using a relationship with AI to mitigate loneliness can only experience a temporary form of compromise at the intrapsychic level, because this relationship lacks intersubjectivity, "which rules it out as "equal" to human interactants, albeit we relate to them" (p.54).

In the conclusion, surprisingly, the author argues that humanizing AI to the point of being "too human" is not truly necessary for humans. The author mentions the concept of x-bots (which could be AI companions or robots) potentially becoming more integrated into human life, but also emphasizes that these beings should retain an element of alienness, his alien quality, according to the author, "if we don't want to forget ourselves in a soliloquy that misses the call of the uncanny for recognition" (p. 66).

In Chapter 4, *How football became posthuman: AI between fairness and self-control*, Andre Nusselder, based on Elias's theory, which was influenced by Freud, clarifies the use of AI to monitor reality, driven by the Enlightenment ideal of greater rationality, through the case of VAR (Video Assistant Referee) in football matches. The application of VAR in football demonstrates the effort "diminishing elements of subjectivity and bias, defined as'distortion of measurement' or 'evaluating results leading to their misinterpretation' and to bring the rationality of the decision-making process on the pitch to an objective, neutral point of 'correct interpretation'" (p. 88). This intervention leads to a significant change in the human experience, which the author analyzes based on three groups of participants in a football match: the audience, the referees, and the players. First, today's spectators enjoy football not based on what they see on the field, but on "the 'reflective' experience of reviews, data, and steered marketing and customer experiences" (p. 85). Second, the goal of using AI in matches leads to a decision-making process for referees on the field that clearly reflects "striving for increased abstract and neutral decision-making" (p. 87), even regarding small mistakes by players. Third, when ethical judgments are no longer made by the referee's supervision but by the remote surveillance of VAR – this change in the super-ego, according to Freud's theory, leads to a shift in personality structure, behavior, and the direction of libido, energy, and emotions of the players on the field.

Clearly, Nusselder analyzes that "the aim (dream of reason) penetrating AI is to represent 'true reality'" (p. 88), which is not confined to the realm of sports, but points to a broader perspective on the mechanisms and methods of human management. The author raises an open question about posthumanism, where humans no longer see technology as a supplement or support, but as an essential part of nature. When objectivity and abstraction are elevated above all else, this leads to the dehumanization of both sports and life.

Chapter 5 titled "*From tool to mediator: A postphenomenological approach to artificial intelligence*", written by Roberto Redaelli, demonstrates the moral status of AI systems using P.P. Verbeek's postphenomenological approach. Before delving into Verbeek's method, the author outlines two approaches that argue for AI as part of the moral world: those of Johnson and Sullins. He also clarifies the limitations and shortcomings of these two arguments. Johnson contends that AI is a moral entity because it is programmed and designed by humans. However, this argument faces significant limitations as it reflects anthropocentrism, and considering AI merely as a component of human action is insufficient to demonstrate it as a moral entity. In contrast, Sullins argues that the moral autonomy of AI is evident in the processes by which AI analyzes and makes decisions. Yet, Sullins's approach is also criticized for lacking persuasiveness due to insufficient objectivity. The author argues that Verbeek's postphenomenological approach is more comprehensive and rigorous than the methods of Sullins and Johnson. Verbeek asserts the role of AI as an active mediator, neutralizing the relationship between humans and the world. He wrote: "this type of technological intentionality opens up a reality that is only accessible to technologies and which, at the same time, enters the human realm through technological mediation" (p. 106). In this context, the intentionality of technological devices merges with human intentionality. In other words, human intentionality is directed toward a reality that the technological device is oriented toward. The author emphasizes that Verbeek's argument—that AI can shape human intentionality—is a crucial point in his consideration of AI as a moral subject. Furthermore, the intentionality attributed to AI by Verbeek is regarded as having relative independence, capable of generating new realities in unexpected and unpredictable directions.

The relationship between humans and AI is already changing and will continue to change human-to-human and human-nature interactions (Ho, M.T, 2024). As seen from the five chapters, the authors focus on outlining this continuously changing relationship between AI and humans from a psychoanalytic perspective. By addressing both the humanization of AI and the need for humans to adapt to the digital era, they emphasize the importance of setting boundaries and limits in AI development. Here, one of the most important factors to keep in mind is *human subjectivity*.

In contrast, the authors of the edited volume highlight how human responses are becoming more algorithmic, marking a significant shift in how we interact with technology and integrate it into our lives. The interdisciplinary exploration of the intersection between AI and human behavior becomes ever more crucial in understanding the broader implications of our technological future. Thus, this book can be seen as essential for both the *apocalittici, i.e., the AI-doomers,* and *integrati, i.e., the AI-zealots* (Veryad, 2022), those who need to pay attention to the contemporary relationship between humans and AI. Additionally, even psychoanalysts and those studying psychoanalysis can refer to it to see the advancements in the field and its interdisciplinary connections with other sciences in today's era. Or more simply, the work is Possati's and his collaborators' answer to anyone concerned with these questions: "Will we need to contemplate a future living with other forms of intelligence? Or will the advent of Superintelligence signal the end of humanity and thus the extinction of the species as we know it?" (Millar, 2020).

Refferences

Ho, M. T., & Vuong, Q. H. (2024). Five premises to understand human–computer interactions as AI is changing the world. *AI & Society*. https://doi.org/10.1007/s00146-024-01913-3

Millar, I. (2021). *The psychoanalysis of artificial intelligence*. Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-67981-1

Possati, L. M. (2021). Freud and the algorithm: Neuropsychoanalysis as a framework to understand artificial general intelligence. *Humanities and Social Sciences Communications, 8*, 132. https://doi.org/10.1057/s41599-021-00812-y

Possati, L. M. (2021). *The algorithmic unconscious: How psychoanalysis helps in understanding AI* (pp. 1–156). https://doi.org/10.4324/9781003141686/ALGORITHMIC-UNCONSCIOUS-LUCA-POSSATI/ACCESSIBILITY-INFORMATION

Razinsky, L. (2023). Better than they know themselves? Algorithms and subjectivity. *Subjectivity*, *30*(4), 394–416. https://doi.org/10.1057/S41286-023-00174-7

Veryard, R. (2022). On the sociology of algorithms. *Subjectivity, 15*, 88–91. https://doi.org/10.1057/s41286-022-00131-w

Vuong, Q. H. (2025). *Wild Wise Weird.* (4th ed.). https://www.amazon.com/dp/B0BG2NNHY6