Chapter 3
Unconscious Artificial Intelligence

Frederic Andres
http://orcid.org/0000-0002-5003-7579
National Institute of Informatics, Japan

ABSTRACT

Unconscious AI encompasses systems that perform intelligent sensing and behavioral analysis in environments populated by entities—such as objects, organisms, or agents—that are incapable of reflecting on their own actions or experiences. These systems themselves do not possess awareness, agency, or understanding, and they function solely to extract and process data from unconscious subjects, without initiating any form of conscious interaction. The lack of self-awareness applies both to the AI system itself as it has no consciousness or self-representation, and to the entities being observed as they are unaware of their own internal state or that they are being observed. This article explores unconscious artificial intelligence.

INTRODUCTION

Artificial intelligence research has historically oscillated between two dominant paradigms: AI as a tool for automated analysis and AI as an intelligent agent capable of autonomous reasoning, planning, and adaptation. In recent years, the latter paradigm—driven by advances in machine learning, reinforcement learning, and large-scale neural architectures—has increasingly shaped both technical development and public discourse. As a result, many AI systems that are fundamentally observational or analytical are implicitly framed using concepts such as goals, intentions, decision-making, or even proto-consciousness, regardless of whether such properties are architecturally present or conceptually warranted.

This chapter argues that this conceptual conflation obscures an important and growing class of systems, which we term Unconscious AI. These systems are engineered to observe, measure, analyze, or model environments composed of entities that themselves lack self-awareness or subjective experience, such as physiological processes, ecological systems, learning traces, or digital signals. Crucially, Unconscious AI systems do not possess self-models, introspective capabilities, intentional stances, or autonomous goals. They neither understand themselves as agents nor represent the entities they monitor as conscious subjects.

DOI: 10.4018/979-8-3693-5638-8.ch003

Copyright ©2027, IGI Global Scientific Publishing. Copying or distributing in print or electronic forms without written permission of IGI Global Scientific Publishingis prohibited. Use of this chapter to train generative artificial intelligence (AI) technologies is expressly prohibited. The publisher reserves all rights to license its use for generative AI training and machine learning model development.

The motivation for introducing Unconscious AI as a distinct category is both theoretical and practical. From a theoretical perspective, debates on artificial consciousness—such as Integrated Information Theory, self-modeling accounts of consciousness, and agent-based AI architectures—require clear negative boundaries: not only what consciousness might require, but also what well-designed AI systems should explicitly exclude. From a practical perspective, many real-world deployments—ranging from educational analytics and environmental monitoring to physiological telemetry—benefit from AI systems that remain strictly non-agentive, non-intrusive, and non-intentional, thereby reducing ethical risks associated with surveillance, manipulation, or emergent autonomy.

Rather than treating the absence of consciousness as an incidental property, this chapter treats it as a design objective. It proposes that Unconscious AI should be deliberately engineered through architectural constraints, functional limitations, and explicit prohibitions against reflective, intentional, or agent-like mechanisms. To this end, the chapter introduces a structured set of design principles, each accompanied by operational rules expressed in first-order predicate logic. This formalization enables Unconscious AI to be specified, evaluated, and audited in a way that is independent of implementation details while remaining robust against unintended conceptual drift.

By defining Unconscious AI as a foundational, non-conscious, and non-agentive paradigm, this chapter seeks to clarify its role within the broader AI ecosystem, distinguish it from human–machine teaming and autonomous systems, and provide designers, researchers, and policymakers with a rigorous framework for building AI systems that are powerful in analysis yet restrained in agency.

WHAT IS AN UNCONSCIOUS AI SYSTEM?

Unconscious AI refers to a class of artificial systems engineered to observe, measure, analyze, or model the behavior of elements within an environment (e.g. physical, biological, or virtual) where the entities being monitored lack any form of self-awareness, introspective capacity, or subjective experience (qualia). These systems can be typically deployed inside telemetry systems to monitor pilot emotion (Frangeto et al., 2021), or to extract behavioral patterns, detect anomalies, support decision-making, or simulate ecosystem dynamics without engaging in reflective cognition or conscious awareness.

This concept extends beyond conventional views of AI as either intelligent agents or goal-driven systems with autonomous capabilities (Russel and Norvig, 2021). In contrast to AI systems that might incorporate symbolic reasoning, goal adaptation or higher-order modeling of self and others such as (Wooldridge, 2002), Unconscious AI systems operate in a purely non-reflective mode. Such systems do not engage in meta-cognition, moral reasoning, or autonomous deliberation. Similarly, the entities it observes—such as human, biological organisms, vegetalble, trees, or non-intentional digital processes—are not aware of their own ongoing states nor of the act of being observed.

The structural role of Unconscious AI within such ecosystems is functional and observational. It processes data streams from sensors, logs, or digital environments and applies statistical or machine-learned models to detect regularities, without representing itself as an actor or modeling the intentions or awareness of the observed subjects. In this regard, it resembles what Sloman and Chrisley call *reactive or deliberative agents*, which operate without entering the reflective layer of cognitive architecture (Sloman & Chrisley, 2003) .

The concept is closely related to debates in the philosophy of mind and artificial consciousness. Tononi's Integrated Information Theory (IIT) has posited that con-

12 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/unconscious-artificial-intelligence/403118

Related Content

Nanomaterials Discovery Using AI-Based Technology Accelerators: Innovation Through Predictive Modeling and Intelligent Experimentation
Ramadevi Vitthal Salunkhe, Tusharkumar Rajaram Sathe, Rushikesh Moreshwarrao Shete, Pooja Wagh, Roshan Kolteand Renuka Raut (2026). *Leveraging AI and Nanotechnology for Materials, Devices, and Manufacturing (pp. 145-190).*
www.irma-international.org/chapter/nanomaterials-discovery-using-ai-based-technology-accelerators/394832

Artificial Intelligence in Cataract Diagnosis and Management With Its Future Directions
N. Ramyaand D. Hemavathi (2025). *Responsible AI for Digital Health and Medical Analytics (pp. 189-210).*
www.irma-international.org/chapter/artificial-intelligence-in-cataract-diagnosis-and-management-with-its-future-directions/365992

A Rule-Based Approach to Automatic Service Composition
Maria J. Santofimia, Xavier del Toro, Felix J. Villanueva, Jesus Barba, Francisco Moyaand Juan Carlos Lopez (2012). *International Journal of Ambient Computing and Intelligence (pp. 16-28).*
www.irma-international.org/article/rule-based-approach-automatic-service/64188

Unsupervised Keyword Extraction Methods Based on a Word Graph Network
Hongbin Wang, Jingzhen Ye, Zhengtao Yu, Jian Wangand Cunli Mao (2020). *International Journal of Ambient Computing and Intelligence (pp. 68-79).*
www.irma-international.org/article/unsupervised-keyword-extraction-methods-based-on-a-word-graph-network/250851

Online Surveillance of IoT Agents in Smart Cities Using Deep Reinforcement Learning
Ahmad Alenezi (2024). *International Journal of Intelligent Information Technologies (pp. 1-15).*
www.irma-international.org/article/online-surveillance-of-iot-agents-in-smart-cities-using-deep-reinforcement-learning/349942