**CS 441/541: Artificial Intelligence, Winter 2023**

**Research Paper Analysis**

**On the ethics of constructing conscious AI**

**Shimon Edelman**

[**Article**](https://arxiv.org/pdf/2303.07439.pdf)

**Chitradevi Maruthavanan**

**PSU ID: 950828319**

In this paper, the author discussed the ethics of constructing conscious AI. The application of ethics to artificial intelligence has evolved from science fiction to a practical challenge in moral philosophy and engineering.AI ethics is dominated by the fear of robots behaving badly towards people, which is reflected in literature. The new discipline of AI ethics is focused on preventing robots from harming humans. The ethics of AI theorists rarely consider the possibility of robots needing protection from their creators. The main goal of AI engineers is to create a tool for human use rather than a companion or peer. The consciousness of robots can lead to suffering, and the blame for that rests on their designers and constructors. Artificial consciousness, like artificial intelligence, must be examined from an ethical standpoint. Negative valence and aversive states can lead to suffering in conscious AI systems. There is an ethical argument against the creation of systems equipped with "synthetic phenomenology". The problems associated with artificial consciousness reflect on the human condition.

The Principle of Pathocentrism holds that only sentient beings have moral standing because they have rights and interests that must be considered. Modern capitalism is likened to a form of slavery because people are renting out themselves. The notion that freedom can be bestowed upon others is criticized for implying that their freedom belongs to us. Pain has a functional purpose in motivating a system to act, according to evolutionary origins and functional roles of pain and suffering in natural sentient systems. The potential emergence of artificial suffering through the appeal of pain as a motivating factor in AI systems is a concern due to the predictive processing and computational framework used to create them.

This paper examines the question of whether sufficiently effective learning and control, with generally good behavioral outcomes, can be achieved in conscious AI systems without suffering. A theory of phenomenality must be a theory of affect, as phenomenal states typically incorporate affective dimensions, especially negative affect, which is the primary component of suffering. Suffering is defined as a state of negative affect from which the sufferer cannot escape by simply wishing it away. The inescapability of suffering is connected to its evolutionary-functional role. The focus of the article is on suffering as it presents itself to the sufferer, rather than ethical problems created for others. Agarwal and Edelman (2020) propose a strategy for avoiding suffering as a direct experience in conscious AI systems by focusing on the phenomenal self-model (PSM), which is an instrument for global self-control and is fundamental to the phenomenology of suffering characterized by a loss of control in addition to negative valence.

The ethical concerns surrounding creating conscious artificial intelligence (AI) and the potential functional benefits of endowing AI with consciousness. The article states that building AI without consciousness is ethically sound as it avoids the risk of suffering for the AI. However, if consciousness confers significant functional advantages for AI systems, then building non-conscious AI may not be viable in commercial settings. The article discusses the benefits of consciousness, such as autonomous self-motivation and effective learning from successes and failures. To mitigate suffering, the article outlines four theoretical options, namely eliminating the PSM, eliminating the NV-condition, eliminating the T-condition, or maximizing the unit of identification. The fourth approach, maximizing the unit of identification, is argued to satisfy the functional needs of consciousness while mitigating suffering.

Agarwal and Edelman argue that shifting a conscious system’s unit of identification away from a self-model would abolish suffering while preserving the functional effectiveness of the modified state of consciousness. Metzinger's conception of suffering is that it presupposes egoic self-awareness. Agarwal and Edelman argue that proper functioning relies on automatic, subpersonal but conscious processes, which can continue unhindered while the system identifies with the MPE upon which these conscious experiences are necessarily superimposed. The move to separate consciousness from suffering without altering the phenomenal nature of the resulting state and its functional effectiveness is difficult and perhaps impossible. The difficulty in separating consciousness from suffering justifies the moratorium on developing conscious AI, proposed by Metzinger. Separating consciousness from suffering is also a key argument for anti-natalism, which sees being born as the greatest evil that can befall a person. Non-egoic systems are not good at dealing with crisis situations or at learning.

The article presents different perspectives and raises ethical concerns about constructing artificial systems that may experience pain through predictive processing. It highlights the need for careful consideration of the ethical implications of such systems and encourages further exploration of the nature of suffering and its place in theoretical treatments of consciousness.