"Conference: Standardization of AI Ethics: Stakeholders, Values, and Profit"

To Be Loved or Deceived? Operationalizing Ethics in the Case of Social Robots Under the European AI Act.

In a world where emerging technologies such as AI and robotics are rapidly expanding and playing a crucial geopolitical and economic role, the European Artificial Intelligence Act (European Commission, 2024) represents one of the first attempts to regulate a field that has typically developed in a legislative vacuum.

This effort is particularly significant given that our relationship with technology is not occasional but an integral part of daily life. Smartphones can almost be seen as extensions of our bodies, and AI-powered devices are becoming embedded in our cognitive, emotional, interpersonal, cultural, and even affective lives. One growing concern in this context is the role of social robotics. Social robots are "a new class of machines designed to function as 'social partners' for humans" (Damiano & Dumouchel, 2020). Care and companionship robots, in example, are created to take on caregiving roles for vulnerable populations, such as the elderly or children with disabilities.

A common feature integrated into robotic design to foster acceptance and long-term use is the simulation of human interaction and characteristics, ranging from linguistic capabilities to aesthetics and behavioural cues. This has led some scholars to raise concerns about authenticity, reciprocity, and deception in human-robot relationships (Sparrow & Sparrow, 2006, Sharkey & Sharkey, 2012)

The ability of social robots to imitate human behaviour and foster attachment feelings in their users could, in fact, be considered manipulative, raising ethical questions about their role in society.

This risk is addressed in Article 5, Chapter II, of the AI Act on Prohibited AI Practices. The article states that "*subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques*" should be prohibited and condemns "*the placing on the market, the putting into service or the use of an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons*" (European Commission, 2024).

In light of this, a crucial question arises: are social robots inherently deceptive, and should they therefore be banned altogether? Should their development be restricted?

Fortunately, philosophy offers valuable insights in two key ways. My argument integrates both top-down and bottom-up approaches, combining ethical and philosophical analysis with the development of an operational framework aimed at informing regulatory policies.

Starting with the theoretical aspect, I will draw on the frameworks of Human-Robot Affective Coordination (Dumouchel & Damiano, 2017) to demonstrate that the common accusation of deception is based on false premises.

In fact, modern philosophy of mind, while affirming the overcoming of Cartesian dualism and recognizing the mind as a cognitive machine like any other, continues to assume the human mind as a paradigmatic epistemic agent. The argument of deception is precisely based on this residual dichotomy, as it demands an authentic correspondence between a personal mind and its external expression. In reality, this is not the case: the mind, as well as emotions and affective processes, exist within a relational space.

This also supported by empirical evidence. Experiments from the behavioural sciences have shown that affective and cognitive processes are bidirectional, with external expressions influencing internal emotional perceptions (Strack et al, 1988, Laird & Lacasse, 2014, Dutton & Aron, 1989) - aligning with the theories of pragmatist philosopher and psychologist William James (James, 1890, Murphy, 1997). Moreover, insights from behavioural and moral psychology suggest that what we describe as introspection is frequently a post-hoc reconstruction (Bem, 1972, Haidt, 2001).

However, specific guidelines must be established, based on an affective and behavioural coordination agreement between humans and robots, a notion grounded in the pragmatist philosophical tradition (Murphy, 1997).

Given this, I will explore how these ethical considerations can be operationalized within regulatory and legal frameworks.

For instance, when it comes to the automation of ethical principles in technological development, several approaches can be considered. One such approach is *Value-Sensitive Design* (Friedman and Kahn 2003), which aims to integrate moral values directly into the design of technological systems. This perspective aligns with *Social Construction of Technology (SCOT)*, which acknowledges that technoscience does not evolve in a deterministic vacuum but is always embedded within social values and practices, and *User-Centred Design*, according to which end-user preferences are crucial in determining the social acceptance of technology. In the specific domain of robotic care, the *Care-Centered Value-Sensitive Design* approach (van Wynsberghe, 2020) proposes the integration of care-related values — such as ensuring eye contact while assisting an elderly person — into the design of care robots. These approaches furnish practical instruments and guidelines to turn theoretical principles into practical regulatory policies.

In conclusion, I will discuss how ethical conclusions regarding human-robot attachment can be translated into regulatory frameworks for the construction, commercialization, and use of social robots to mitigate the risk of deception.

References

Bem, D. J. (1972). Self-perception theory. *Advances in experimental social psychology*, *6*.

Damiano, L., & Dumouchel, P. G. (2020). Emotions in relation. Epistemological and ethical scaffolding for mixed human-robot social ecologies. *HUMANA. MENTE Journal of Philosophical Studies*, *13*(37), 181-206.

Dumouchel, P. & Damiano, L. (2017). Living with robots. Harvard University Press.

Dutton, D. G., & Aron, A. (1989). Romantic attraction and generalized liking for others who are sources of conflict-based arousal. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *21*(3), 246.

European Commission (2024), Artificial Intelligence Act (Regulation (EU) 2024/1689), Chapter II, Article 5, Prohibited AI Practices, Official Journal version of 13 June 2024.

Friedman, B., & Kahn, P. H., Jr. (2003). Human values, ethics, and design. In The human–computer interaction handbook: Fundamentals, evolving technologies and emerging applications (pp. 1177–1201). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

James, W. (1890). The principles of psychology. *Henry Holt*.

Laird, J. D., & Lacasse, K. (2014). Bodily influences on emotional feelings: Accumulating evidence and extensions of William James's theory of emotion. *Emotion Review*, *6*(1), 27-34.

Murphy, JP. (1997). Il pragmatismo. Il mulino.

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and information technology*, *14*, 27-40.

Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. Minds and Machines, 16, 141-161.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. Journal of Personality and Social Psychology, 54(5), 768-777.

Van Wynsberghe, A. (2020). Designing robots for care: Care centered value-sensitive design. In *Machine ethics and robot ethics* (pp. 185-211). Routledge.