Reimagining Artificial Intelligence through the Critical Lens of Harry Frankfurt

Neeraj Sharma

Master of Arts in Philosophy, Indira Gandhi National Open University

Abstract

Harry Frankfurt's philosophy of "free will" presents a significant departure from classical compatibilist, in-compatibilist, and libertarian accounts of free will by shifting the focus from alternative possibilities to the structure of human agency. Frankfurt argues that what matters for a person to have free will1 is not the ability to do otherwise, but the capacity of reflective endorsement of one's desires, a hierarchical model of the will. He distinguishes first-order desires from second order desires, suggesting that, in order to possess free will an agent must have alignment between these levels of desire and its volition. This reflective structure of Frankfurt philosophy is central to the concept of personhood and free will. Frankfurt defines personhood in terms of capacity of an agent to have second-order volitions, especially the ability of reflective self-endorse of desires.

Frankfurt's conception of personhood is framed in psychological and volitional terms, not strictly biological or species-based. So, in principle any being biological or artificial that possesses second-order volitions can qualify to be a person in Frankfurt's sense. In the philosophy of mind, the prominent theory of "functionalism" holds that mental states are defined by their functional roles, how they interact with and what role they play, not what they constitute of. This "functional" and "hierarchical" conception, theoretically, allows for non-human persons, including AI, to possess intelligence, cognition and free will, if they have the right kind of structure, what matters for intelligence or personhood is the pattern of operations, not the material from which the being is made.

This paper explores the possibility of attributes such as mental states, cognition, free will and intelligence, which are exclusive to Human beings, in Artificial Intelligence machines and critically evaluates the functionalist view of mind with reference to Frankfurt's conception of personhood and free will. This paper aims to initiate an inquiry into the extent to which attributing features of human cognition to artificial intelligence is justified.

Keywords: - Artificial Intelligence, Concept of person, Free will, Second-order volition

Introduction

Harry Frankfurt's theory of free will and his conception of personhood, provides a new approach to the philosophical discourse surrounding artificial intelligence, particularly regarding the question of whether AI can attain or appropriate the essential features of human personhood. In Frankfurt philosophy, problem of free will has occupied a central place, particularly within the domains of moral responsibility and personal identity. Harry Frankfurt, in his paper, "Freedom of the Will and the Concept of a Person" (1971), offer a nuanced and psychologically grounded theory of free will. He reorienting the debate from traditional metaphysical dichotomy of determinism versus indeterminism to the structure of the will and introduces a compelling account of personhood grounded in second-order volitions. This paper will explores Frankfurt's theory of free will and his concept of personhood, analyses how his ideas reconfigure long-standing philosophical debates.

Frankfurt's view on Free Will and Moral Responsibilities

Frankfurt critiques the classical framing of the free will problem. Traditionally, the debate hinges on whether human freedom can exist in a determined universe which runs by rigid Natural Laws. Compatibilists attempt to reconcile freedom with determinism. In-compatibilist opposes the idea, whereas Libertarians argue for an indeterministic form of free will. While Frankfurt believes that the real issue lies not in metaphysical speculation but in the structure of the will and the psychological capacities of agents. For Frankfurt, what matters is whether a person's will is aligned with their deeper values, not whether their choices are uncaused.

A cornerstone of Frankfurt's theory is his hierarchical model of the will, which distinguishes between different levels of desires; First-order desires, second order desires and volition of second order desires. First order desire are simple wants directed toward actions or outcomes. For instance, the desire to eat cake or to stay in bed are first-order desires. Second-order desires are desires about desires. That is, one may desire to have or not to have a certain first-order desire. For example, a dieter might have a first-order desire to eat cake but a second-order desire not to want cake. The critical distinction lies in second-order volitions which occurs when an individual not only has a second-order desire but also wants that desire to be effective in determining their will. In Frankfurt's terms, "a person wants a certain desire to be their will", i.e., they want it to be the desire that leads to action.

Frankfurt challenges the Principle of Alternate Possibilities (PAP) according to which a person is morally responsible for their actions only if they could have done otherwise. Frankfurt in his famous thought experiment constructed a scenarios in which an agent appears to be morally responsible despite lacking alternative options. He devised a situation where an agent decides to perform an action let's say voting for a candidate, an external mechanism was in placed in his brain to force the action only if the agent had chosen otherwise. Because the agent chose freely, even though they couldn't have done otherwise. Frankfurt argues that they are still morally responsible. Through this thought experiment Frankfurt proofs that moral responsibility depends on the internal structure of the will, not on the availability of alternative courses of action.

The Concept of a Person

Persons, according to Frankfurt's framework, are beings who are capable of self- reflecting on their desires and forming volitions. Frankfurt's definition of personhood is one of the most distinctive elements of Frankfurt's theory. He argues that what distinguishes persons from non-persons is not the biological traits but the capacity of forming a second-order volitions, this feature is absent in most of the biologically similar species. Wantons, a term Frankfurt introduces, are beings that have desires to act, but lack second-order volitions. They do not care which desires move them to act. They have freedom of actions but not free will neither they are considered morally responsible. A human being who lives entirely in the moment, never reflecting on or endorsing their motives, could also be a wanton. Frankfurt contends that the capacity to form second-order volitions is a necessary condition for being morally responsible. A person can be held morally accountable for their actions when those actions flow from a will they endorse. This necessarily implies that one can be morally responsible even in a deterministic universe because what matters is not the presence of causal determinacy or alternate possibilities for an agent but the authenticity of one's volitions.

AI and the Personhood

We are living in the age of Artificial intelligence recent decades have shown an exponential growth in artificial intelligence. The distinction between men and machine are blurring. Now a days, it seems that only the biological difference left alone to distinguish human intelligence from machine intelligence. Machines grow increasing capacity of performing tasks that were once thought to be exclusive to human faculties, example; learning, reasoning, decision-making etc. the notion of intelligence becoming blurred. These development poses a critical challenges for the philosophers that whether the general capacities that define a person and his personhood, such as rational, free will, moral responsibility, and mental states are still exclusive to human, or they can be instantiated in artificial systems in near future?

The functionalist theory of mind, which holds that mental states are defined by their causal and functional roles rather than their physical substrate, allows the possibility of extension of mental states and thereby the personhood to be realized in machines.

Now we will analyse, how Frankfurt's theory, in conjunction with functionalism, contributes to the contemporary debate of whether AI can possess mental states and personhood or we will continue with our anthropocentric bias towards AI? Can non-human entities meet the philosophical criteria for being persons? Whether or not AI systems can have "minds" in a meaningful sense? By combining conceptual analysis and philosophical argumentation, I aims to explore the assumptions that moral and mental capacities are biologically exclusive and whether AI could meaningfully participate in the realm of moral agency and personhood?

Frankfurt's introduction of hierarchical model of desires, says that all agents have first-order desires, which are simple urges or wants to perform action but only a persons can form second-order desires. What differentiates a person from non-person is the capacity to form second-order volitions that is, not merely to have second-order desires but to will that a certain first-order desire be the one that moves them to action. This form of self-reflective governance is essential for an agent to be morally responsible. Frankfurt does not give emphasis to the physical substrate in which this complexity is realized, it could be in biological and non-biological. Nothing in Frankfurt's theory inherently excludes the possibility that non-human entities could not possess the required structure of the will. In principle, it could be said that if an AI system is representing, evaluating, and endorsing its own motivational states, then, it would not be wrong in appropriating second-order volitions to artificially intelligence system and consequently appropriating personhood to AI.

This model allows more inclusive understanding of personhood one that is not confined to biological or human entities. It opens the door for artificial agents, given sufficient sophistication, to qualify as moral agent. This reinterpretation will have profound implications for how we might assign moral status or responsibility to AI systems and whether we might one day view them not merely as tools but as participants in ethical life.

Functionalist defines mental states not by what they are made of but by what causal role they play to sensory inputs, behavioral outputs, and other mental states. A mental state like pain, for instance, is not necessarily a specific neural event but a state that causes avoidance of certain behavior, distress, and is triggered by injury. This allows a theoretical possibility that mental states could be realized in radically different kinds of systems in which mental life is substrate-independent i.e what matters is not its biological composition but the system's architecture of processing and internal representation of cause and then producing output.

Functionalism provides the most robust theoretical framework for understanding how machines could possess minds. A sufficiently advanced AI system which can processes inputs, selects outputs based on a dynamic network of desired goals and rules and forming some internal representation could be, in principle, said to have mental states. If such a system can reflect on its internal states, revising its goals, and modifying its behaviors accordingly, it may exhibit the kind of self-governing functional architecture that functionalists associate with consciousness and cognition. The integration of functionalist and Frankfurt models thus sets the conceptual groundwork for a more inclusive theory of personhood in which not only person but also AI can instantiate the relevant structures of agency, reflection, and identification.

To support my theory, I devised a hypothetical argument, imagine an AI system, Brahman, a next-generation AI model designed to learn from social interaction and ethical dilemmas, it has access to a deep neural architecture, it can morally evaluate, and learn from its consequences. It goes beyond basic decision-making. Brahman is programmed not only to revise its goals but also to monitor and evaluate the motivational structure behind those goals. Over time, Brahman developed itself and behave exactly in way as humans do.

Now, according to Frankfurt's framework, second-order volitions is the essential marker of personhood in humans, this involves reflective deliberation, complex cognitive and emotional feedback systems, a sense of moral identity and under environmental influence behaviour. Now the question arises, can these components be functionally realized in machines? Functionalism affirms positively. Now, if second-order volitions are defined by their causal-functional roles such as identifying with certain motivations which moved the agent into action, suppressing conflicting urges, and shaping future behaviour, then it can be plausibly conceivable that AI systems could fulfil these needs of meta-cognition, goal-rearrangement, without being biologically human.

Moreover, recent developments in AI research reinforce the plausibility of this view. Models like OpenAI's GPT, Google's DeepMind systems, and adaptive RL (reinforcement learning) shows that an agents can learn not only from immediate rewards but also from delayed feedback and changing environments. These systems can form internal representations of values, self-correct itself based on performance, and even discuss ethical trade-offs. If not full moral agency, prima facia, it looks that they possess some foundational cognitive architecture.

If such systems is developed then personhood, as Frankfurt defines it, might no longer be exclusive to humans. We might be compelled to treat these systems not merely as sophisticated tools but as agents with a degree of autonomy and moral consideration. The functionalist theory of mind, coupled with Frankfurt's theory of personhood, provides a robust conceptual foundation for extending human-like capacities to AI. If AI systems can be built with the cognitive and volitional complexity required for second-order volitions, then they meet the criteria for personhood not by analogy, but by philosophical and theoretical justification. The line that separates humans from machines would then rest only on biology, not on a shared capacity for rational reflection. The philosophical question of whether AI can possess mental states and qualify as persons is no longer speculative. As artificial systems become more sophisticated, it is imperative to establish clear conceptual frameworks for understanding the conditions under which they may be regarded as agents, moral subjects, or even members of a community of persons.

Critical Evaluation: Rethinking Personhood in the Age of AI

While Frankfurt's account provides a theoretically sound and psychologically plausible model for human moral agency, extending this model to AI involves significant philosophical and practical challenges. Frankfurt's notion of personhood relies on the capacity for reflective endorsement. This identification is not a mechanical preference but an evaluative process involving moral and psychological depth. A major challenge arises when we ask examine the genuineness of second-order volitions in an AI systems.

If an AI is programmed to revise its goals based on meta-level rules or human feedback, its "identification" with a goal may be a product of functional programming rather than genuine self-evaluation. Major criticism is that AI lacks phenomenal consciousness; it does not feel or experience desires. Its evaluations are externally imposed and manufactured rather than autonomously generated. This raises the question: Is it enough to functionally model second-order volitions, or such volitions must be subjectively owned? AI can be programmed to say "I want to want to help humans." It can revise its goals based on higher-level reasoning. But the question arises that is this a reflective endorsement, or simply algorithmic optimization? This brings us to John Searle criticism, who claim that AI can only simulate and cannot replicate the actual human being. His distinction of strong AI and weak AI. The possibility of strong AI is denied by Searle but he supported the possibility of weak AI.

One can argue that Frankfurt's theory is purely structural: if a system exhibits the right hierarchical relationships among motivational states, it counts as a person. This functionalist reading opens the door for AI personhood. However, Frankfurt never divorces volitional structure from the phenomenology of agency. When a person identifies with a desire, it feels right it resonates with their sense of self. AI, devoid of inner experience, may behave like a moral agent, but without the subjective engagement, it cannot fully qualify as one.

Critics argue that AI's intelligence is mere simulation which is devoid of, what they refers is consciousness which is exclusive to human being only. But such critiques are often based anthropocentric biases which rests on ambiguous or dualist assumptions about consciousness. If we logically analyses, human second-order volitions are also causally determined. Our desires and volitions shaped by evolution, neural architect, and social conditioning. These events defines human self yet we regard them as authentic. In this sense if AI reflections are the product of structured causal histories, then the distinction between "real" and "simulated" volition becomes philosophically redundant.

In conclusion, Frankfurt's theory provides a rich framework for understanding human moral agency which is based on capacity for second-order volitions and self-reflective identification. We have also seen that extending this theory to AI reveals multiple layers of difficulty. The present day AI which lacks the phenomenological depth and emotive grounding that give human volitions their moral weight and existential meaning. AI's inability to genuinely reflect, feel, or experience undermines its claim to possess personhood in the Frankfurtian sense. Nevertheless, there are challenges but it do not out rightly closes the debate. It pushes us to further clarify what it means to have volition, consciousness, and personhood and whether new this forms of intelligence which is evolving rapidly may demand new philosophical categories which are more exclusive and distinct. As AI continues to evolve, we may need a revised or expanded our Frankfurtian framework of personhood one that balances the structural features of agency with the elusive but essential qualities of inner life.

References

1. Frankfurt, Harry G. (1971). Freedom of the Will and the Concept of a Person. Journal of Philosophy, 68(1), 5–20.
2. Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. The Journal of Philosophy, 68(1), 5–20. https://doi.org/10.2307/2024717
3. Frankfurt, Harry G. (1988). The Importance of What We Care About. Cambridge University Press.
4. Putnam, Hilary. (1967). Psychological Predicates. In W.H. Capitan & D.D. Merrill (Eds.), Art, Mind, and Religion.
5. Block, Ned. (1978). Troubles with Functionalism. In Readings in Philosophy of Psychology, Vol. 1.
6. Searle, John R. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences, 3(3), 417–457.
7. Penrose, Roger. (1989). The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics. Oxford University Press.
8. Warwick, Kevin. (2003). Cyborg Morals, Cyborg Values, Cyborg Ethics. Ethics and Information Technology, 5(3), 131–137.
9. Turing, Alan M. (1950). Computing Machinery and Intelligence. Mind, 59(236), 433–460.
10. Dennett, Daniel C. (1991). Consciousness Explained. Little, Brown and Co. – A functionalist view that strongly supports cognitive modeling in AI.
11. Nagel, T. (1974). What Is It Like to Be a Bat? The Philosophical Review, 83(4), 435–450. https://doi.org/10.2307/2183914