

AI and Ultimate Responsibility: Reinterpreting Robert Kane's Theory of Moral Responsibility

Zahra Khazaei 

Professor, Department of Philosophy, University of Qom, Qom, Iran. E-mail: z-khazaei@qom.ac.ir

Article Info

Article type:

Research Article

Article history:

Received 02 July 2026

Received in revised form

01 August 2026

Accepted 18 August 2026

Published online 25

October 2026

Keywords:

Artificial Intelligence;
Moral Responsibility;
Robert Kane; Ultimate
Responsibility; Self-
Forming Actions; Human
Agency; AI Ethics

ABSTRACT

The increasing integration of artificial intelligence (AI) into human decision-making has raised questions about the adequacy of theories of moral responsibility. This article examines whether Robert Kane's libertarian theory remains an adequate framework for explaining moral responsibility under AI-mediated conditions. Using an analytical–philosophical method, the study evaluates Kane's central concepts of Ultimate Responsibility, Self-Forming Actions, and reflective control. The analysis argues that AI does not eliminate responsibility because Kane's account depends on autonomous agency rather than the absence of technological assistance. However, AI transforms the conditions under which responsibility is exercised and requires reinterpretation of these concepts within technologically mediated environments. The study argues that decisions concerning reliance on, evaluation of, and oversight over AI systems may constitute contemporary forms of Self-Forming Actions because they involve reflective judgments contributing to moral self-formation. Human agents remain the ultimate bearers of responsibility as long as they retain the capacity for critical deliberation, endorsement, rejection, and revision of algorithmic recommendations. Accordingly, this article argues that if future AI systems were to undermine these fundamental conditions of human agency, the resulting challenge would extend beyond Kane's libertarian framework to broader theories of moral responsibility. The article concludes that Kane's framework remains valuable for understanding responsibility in the age of artificial intelligence.

Cite this article: Khazaei, Z. (2026). AI and Ultimate Responsibility: Reinterpreting Robert Kane's Theory of Moral Responsibility. *Journal of Philosophical Investigations*, 20(56), 151-174. <https://doi.org/10.22034/jpiut.2026.73339.4608>



© The Author(s).

Publisher: University of Tabriz.

Introduction

The relationship between agency, free will, and moral responsibility has long been regarded as one of the most fundamental problems in moral philosophy and the philosophy of action. The emergence of artificial intelligence (AI) systems has introduced this classical philosophical problem into a new stage, transforming it from a predominantly theoretical debate into an interdisciplinary issue with significant practical implications. Consequently, the central question is no longer merely how human beings can be held morally responsible for their actions, but rather on what grounds moral responsibility can be attributed to human agents when decisions are made in interaction with intelligent systems.

The widespread adoption of AI in healthcare, judicial decision-making, transportation, finance, security, and public administration has fundamentally altered ethical decision-making. Rather than resulting exclusively from direct human deliberation, moral decisions increasingly emerge through interactions between human agents and algorithmic systems, generating a growing interdisciplinary literature on agency, free will, and moral responsibility (Floridi & Sanders, 2004; Matthias, 2004; Vallor, 2016; Coeckelbergh, 2020).

One of the central issues within this literature is the so-called *Responsibility Gap*. As AI systems acquire greater operational autonomy, determining who deserves moral praise or blame for the consequences of AI-assisted decisions becomes increasingly difficult. Some scholars argue that this situation exposes the inadequacy of traditional theories of moral responsibility, which were primarily developed for contexts of direct human decision-making and therefore lack the conceptual resources to explain the complex relationships among human agents, algorithms, and technological systems (Matthias, 2004). Others maintain that although decision-making structures have changed, responsibility itself remains fundamentally human; what requires reconsideration is the explanation and attribution of responsibility rather than its existence (Floridi, 2013; Coeckelbergh, 2023).

Among contemporary theories of moral responsibility, Robert Kane's account occupies a particularly significant position. Defending a libertarian conception of free will, Kane grounds moral responsibility in three interrelated concepts: Ultimate Responsibility (UR), Self-Forming Actions (SFAs), and reflective control. Kane holds that individuals can be regarded as genuinely morally responsible only if they are the ultimate sources of both their actions and the formation of their moral character. Accordingly, moral responsibility is grounded not merely in causal attribution or psychological control, but in the agent's active participation in shaping his or her moral character through practical deliberation (Kane, 1996; Kane, 2005).

In his later reformulation of libertarianism, Kane reaffirms that the primary philosophical significance of free will lies in preserving Ultimate Responsibility rather than merely securing alternative possibilities. Responding to recent criticisms, he emphasizes that self-forming

choices are significant because they open genuinely new pathways through which rational agents shape their future moral identity. As Kane states, "Every undetermined self-forming choice is the initiation of a novel pathway into the future, whose justification lies partly in that future and is not fully explained by the past." (Kane, 2021, 23).

Despite its philosophical significance, Kane's theory has remained the subject of continuing debate. Critics have questioned whether libertarian indeterminism can adequately explain moral responsibility without succumbing to the problem of luck. Alfred Mele and Ishtiyaque Haji have argued that indeterministic choice may weaken rather than strengthen genuine agent control (Mele, 2006). More recently, Neil Campbell has questioned whether Kane's later reformulation preserves the symmetry of efforts of will that characterized his earlier account, arguing that this development raises important questions concerning the continuity of Kane's theory (Campbell, 2023). Nevertheless, these criticisms have reinforced rather than diminished the central place of Kane's theory in contemporary debates on free will and moral responsibility.

The present study does not seek to defend Kane's theory uncritically, nor does it reject it on the basis of these criticisms. Rather, it proceeds from the assumption that the emergence of artificial intelligence creates a new philosophical context in which both the strengths and the limitations of Kane's theory become more visible. Whereas previous debates have focused primarily on determinism, alternative possibilities, or luck, AI-assisted decision-making introduces a different challenge: whether a theory originally developed to explain the moral responsibility of individual human agents can adequately account for responsibility within technologically mediated environments involving algorithms, designers, institutions, and users.

Accordingly, rather than simply applying Kane's theory to artificial intelligence, this article critically examines whether its central concept, Ultimate Responsibility, Self-Forming Actions, and reflective control, require reinterpretation to extend and preserve their explanatory power in technologically mediated decision-making. The central research question of this article is therefore whether Robert Kane's theory of moral responsibility continues to provide an adequate account of human moral responsibility in the age of artificial intelligence and, if so, what conceptual reinterpretations are required to preserve its explanatory power.

The guiding hypothesis advanced in this article is that the normative foundations of Kane's theory retain substantial explanatory value. Nevertheless, their application to AI-assisted decision-making requires a reinterpretation of several of the theory's key concepts. Specifically, this study argues that Ultimate Responsibility should be understood within contemporary structures of technologically mediated decision-making; Self-Forming Actions should be reconsidered in the context of human beings' critical engagement with intelligent technologies; and the concept of control should be reconceptualized in terms of human oversight of algorithmic processes together with the preservation of the guiding role of human agency. From

this perspective, artificial intelligence should be understood not as the end of human moral responsibility but as an occasion for rethinking the conditions under which moral responsibility is exercised and attributed.

Research Method

This study is a theoretical inquiry in moral philosophy employing an analytical-critical methodology. The research begins by reconstructing Kane's fundamental concepts through close analysis of his principal works, particularly *The Significance of Free Will* (1996), *A Contemporary Introduction to Free Will* (2005), and his subsequent publications. These concepts are then examined in light of contemporary debates on AI ethics, artificial agency, and moral responsibility.

The methodology is not merely descriptive but fundamentally critical. Following systematic reconstruction, Kane's explanatory capacity is evaluated regarding AI challenges including increasing autonomy, algorithmic decision-making, the Responsibility Gap, and transformation of traditional control models.

To this end, the study critically examines Kane's theory in light of major contemporary critiques of moral responsibility (Fischer & Ravizza, Mele, Pereboom, and Clarke) and recent contributions to AI ethics (Matthias, Floridi & Sanders, Coeckelbergh, Vallor, and Gunkel) in order to evaluate both the explanatory power and the limitations of Kane's account in AI-mediated decision-making

The argument is organized in three stages: first, reconstruction of Kane's conceptual foundations; second, analysis of principal AI challenges to moral responsibility; third, critical evaluation of Kane's explanatory potential and proposal of a reconstructed interpretation for technologically mediated decision-making.

Literature Review and Theoretical Framework

The question of moral responsibility has long been one of the central concerns of moral philosophy and the philosophy of action. However, the rapid development of artificial intelligence and the increasing integration of intelligent systems into various aspects of human life have extended this issue beyond the traditional debates on free will and determinism, transforming it into one of the most significant topics in the philosophy of technology and applied ethics. Accordingly, the literature relevant to the present study has evolved along two relatively independent yet closely interconnected lines of inquiry: first, research on theories of free will and moral responsibility; and second, studies concerning AI ethics and responsibility in intelligent systems. The purpose of this section is not merely to review previous scholarship but to clarify the theoretical position of the present research within these two intellectual traditions.

Within the literature on moral responsibility, a considerable body of research has been devoted to identifying the conditions under which moral responsibility can legitimately be attributed and to examining its relationship with free will. Among contemporary theories, Robert Kane's libertarian account of free will represents one of the most influential attempts to defend moral responsibility against determinism. By introducing concepts such as Ultimate Responsibility (UR) and Self-Forming Actions (SFAs), Kane argues that moral responsibility can be legitimately attributed only when individuals play a decisive role in the formation of their own will and moral character (Kane, 1996, 2005). The significance of Kane's theory lies in the fact that freedom is understood not simply as the ability to choose among alternative courses of action, but rather as a developmental process through which an individual's moral character is gradually constituted by fundamental decisions.

Despite its considerable influence, Kane's theory has been the subject of extensive philosophical criticism. John Martin Fischer and Mark Ravizza, defending their influential account of guidance control, argue that moral responsibility does not necessarily require the availability of alternative possibilities. Instead, what matters is the agent's possession of an appropriate form of rational control over action (Fischer & Ravizza, 1998). Likewise, Alfred Mele, by focusing on the problem of moral luck, questions whether fundamentally indeterministic decisions can genuinely serve as the basis of moral responsibility (Mele, 2006). Derk Pereboom, from the perspective of hard incompatibilism, casts doubt on the very possibility of moral responsibility in its traditional sense (Pereboom, 2014). Furthermore, Randolph Clarke, while defending several libertarian assumptions, raises important concerns regarding the relationship between indeterminism and moral agency (Clarke, 2003).

Among recent contributions, John Lemos has defended a libertarian account closely aligned with Kane's conception of free will (2021), whereas Ishtiyaque Haji has continued to challenge libertarian approaches, particularly regarding the luck objection (2022). Their exchange in the *Journal of Philosophical Theological Research* (Haji, 2023; Lemos, 2023) illustrates that debates over libertarian free will remain active and unresolved.

Kane himself has also refined his position in his later writings. In (2021), he further develops his account by responding to longstanding objections concerning luck and indeterminism. Kane emphasizes that free actions should never be understood as random occurrences; rather, they are decisions in which agents actively participate in shaping their future moral character. Although this later formulation provides a more sophisticated account of libertarian free will, it nevertheless remains situated within the traditional philosophical debates on free will and determinism and does not explicitly address the challenges introduced by artificial intelligence and technologically mediated decision-making.

The second major body of literature relevant to the present study concerns the ethics of artificial intelligence. This rapidly expanding field has primarily focused on the question of

whether the increasing participation of intelligent systems in decision-making processes requires a fundamental revision of traditional theories of moral responsibility. One of the most influential contributions to this debate is Andreas Matthias's notion of the "Responsibility Gap." Matthias argues that, because the behavior of learning systems cannot always be fully predicted by their designers or operators, attributing responsibility to human agents becomes significantly more difficult in AI-mediated decision-making contexts (Matthias, 2004). His account has become one of the most widely discussed frameworks in contemporary debates on AI ethics and has shaped much of the subsequent literature.

In contrast, Luciano Floridi and Jeff Sanders, while acknowledging that intelligent technologies complicate the structure of moral responsibility, maintain that moral responsibility remains fundamentally human. They argue that although artificial systems may perform actions with significant ethical consequences, they should not automatically be regarded as fully fledged moral agents. Their concept of Artificial Moral Agency (AMA) therefore distinguishes between the functional role played by intelligent systems and the normative status required for genuine moral responsibility (Floridi & Sanders, 2004). Floridi later develops this argument within his philosophy of information, emphasizing that intelligent technologies transform the informational environment in which moral decisions are made without thereby replacing the human moral subject (Floridi, 2013).

A similar perspective can be found in the work of Mark Coeckelbergh, who argues that the primary challenge posed by artificial intelligence is not the replacement of human beings by machines, but rather the transformation of moral relationships and patterns of responsibility within technologically mediated environments. According to Coeckelbergh, AI reshapes the contexts in which human responsibility is exercised while leaving human moral agency at the center of ethical evaluation (Coeckelbergh, 2020, 2023).

Likewise, Shannon Vallor, drawing upon virtue ethics, argues that intelligent technologies should not be understood merely as external tools or as substitutes for human judgment. Instead, technologies participate in shaping the moral development of individuals by influencing the cultivation of virtues and practical wisdom. Consequently, the ethical significance of AI lies not in granting machines moral status, but in examining how human beings interact with technological systems and how these interactions affect the formation of moral character (Vallor, 2016).

Other scholars have approached the issue from different perspectives. Wendell Wallach and Colin Allen (2009) examine the possibility of constructing morally sensitive artificial systems while emphasizing the continuing necessity of human ethical oversight. David Gunkel (2012, 2018) questions traditional assumptions concerning moral agency and responsibility by exploring whether intelligent systems may eventually require new categories of ethical consideration. Nevertheless, even these approaches generally distinguish between the

functional capacities of intelligent systems and the richer normative characteristics associated with human moral agency.

More recently, discussions concerning agency have acquired renewed significance. The emergence of artificial intelligence has prompted a careful ontological reconsideration of the concept of agency itself. Although intelligent systems increasingly participate in decision-making processes, their functional effectiveness alone is insufficient to establish genuine moral agency. A clear distinction must therefore be maintained between the operational capacities of intelligent systems and the ontological conditions required for moral agency and normative responsibility. This distinction provides an important philosophical foundation for evaluating responsibility in human–AI interaction and supports the broader claim that moral responsibility ultimately remains grounded in human agency rather than technological functionality (Khazaei, 2026).

Taken together, these studies demonstrate that contemporary AI ethics has generated a rich discussion concerning artificial agency, algorithmic decision-making, responsibility gaps, transparency, explainability, and institutional accountability. At the same time, however, this literature has rarely engaged directly with contemporary libertarian theories of moral responsibility, particularly Robert Kane's account of Ultimate Responsibility and Self-Forming Actions.

Despite the remarkable expansion of both the literature on moral responsibility and the rapidly growing scholarship on AI ethics, these two fields have largely developed independently of one another. Research on Robert Kane's theory has remained largely within the traditional philosophical debates concerning free will, determinism, alternative possibilities, Ultimate Responsibility, and moral luck. Even Kane's most recent contribution, (2021), while offering a more refined formulation of libertarian free will, continues to operate primarily within the classical framework of the free will debate and does not explicitly address the philosophical challenges raised by artificial intelligence.

Conversely, the contemporary literature on AI ethics has concentrated predominantly on questions of artificial agency, autonomous systems, algorithmic decision-making, responsibility gaps, explainability, transparency, accountability, and AI governance. Although these studies have significantly advanced our understanding of ethical issues surrounding intelligent systems, they have rarely examined the explanatory potential of contemporary philosophical theories of moral responsibility, particularly Robert Kane's libertarian account. Consequently, a systematic investigation of Kane's theory in relation to the philosophical challenges posed by artificial intelligence remains largely absent from the existing literature.

This research gap is particularly significant because many of the central problems discussed in AI ethics ultimately concern the philosophical foundations of moral responsibility itself.

Once intelligent systems become participants in decision-making processes, the primary issue is no longer limited to the accuracy or efficiency of algorithmic outputs. Rather, the fundamental philosophical question concerns how moral responsibility should be attributed among human decision-makers, system designers, software developers, institutions, regulatory bodies, and intelligent technologies. Addressing this question requires a careful re-examination of philosophical theories that ground moral responsibility in human agency.

At the same time, a substantial portion of the contemporary literature on AI ethics proceeds from the assumption that the growing complexity and autonomy of intelligent systems necessarily require a fundamental revision of traditional concepts of responsibility. Consequently, many scholars have increasingly focused on theories of distributed, collective, or institutional responsibility, while the continuing relevance of individual accounts of moral responsibility under AI-mediated conditions remains less examined.

While such approaches undoubtedly illuminate important dimensions of AI governance, they also risk diminishing the normative significance of individual agency and the philosophical foundations upon which moral responsibility has traditionally rested.

The present study adopts a different strategy. Rather than abandoning classical theories of moral responsibility in response to emerging technologies, it argues that their explanatory capacity should first be critically reassessed. More specifically, this article investigates whether Robert Kane's theory of moral responsibility, although originally formulated within the traditional debates on free will, still possesses sufficient conceptual resources to explain moral responsibility in the age of artificial intelligence. It further argues that the continuing relevance of Kane's theory depends not upon preserving its concepts unchanged, but upon reinterpreting several of its central notions in light of technologically mediated decision-making.

Accordingly, the originality of the present research does not lie merely in presenting another interpretation of Kane's theory, nor simply in contributing to the literature on AI ethics. Rather, its principal contribution consists in establishing a philosophical dialogue between two fields that have rarely been examined together. Instead of asking whether artificial intelligence has eliminated human moral responsibility, this study seeks to demonstrate how Kane's central concepts, Ultimate Responsibility, Self-Forming Actions, and human agency, may be reconstructed so as to account for the new conditions created by intelligent technologies.

The theoretical framework adopted throughout this article therefore rests upon the assumption that artificial intelligence neither replaces the human moral agent nor renders the foundations of moral responsibility obsolete. Instead, intelligent technologies transform the conditions under which responsibility is exercised, attributed, and evaluated. Consequently, the central philosophical question is not whether moral responsibility survives the emergence of AI, but rather whether Kane's conceptual framework possesses sufficient explanatory power to

account for these transformed conditions and, where necessary, which elements of his theory require philosophical reconstruction.

On this basis, the following sections reconstruct the central concepts of Robert Kane's theory of moral responsibility and critically examine their applicability to contemporary AI-mediated decision-making. Through this analysis, the article proposes a revised interpretation of Kane's framework that preserves the normative foundations of human moral responsibility while accommodating the philosophical challenges introduced by artificial intelligence.

The Foundations of Robert Kane's Theory of Moral Responsibility

Robert Kane's theory of moral responsibility is widely recognized as one of the most influential contemporary accounts of libertarianism. Among contemporary interpretations of libertarian free will, Kane's account is widely regarded as distinctive because it places Ultimate Responsibility and Self-Forming Actions at the center of moral responsibility (Palmer, 2014).

Its significance extends beyond defending free will to providing a systematic account of moral responsibility that integrates freedom, agency, and moral desert. Kane argues that individuals are fully morally responsible only when they are not merely the immediate causes of their actions but also the ultimate authors of the will and moral character from which those actions arise (Kane, 1996, 28–34; Kane, 2005, 120–138).

Kane develops this account in response to two influential traditions in the free will debate. Hard determinists deny that genuine moral responsibility is possible because every action is causally determined and individuals could never have acted otherwise (Pereboom, 2014). Compatibilists, by contrast, argue that moral responsibility is compatible with determinism, provided that agents exercise an appropriate form of rational control over their conduct (Fischer & Ravizza, 1998). Kane rejects both positions. In his view, rational control alone cannot ground genuine moral responsibility if individuals have played no significant role in forming the values, commitments, and character from which their actions originate (Kane, 1996).

The distinctive feature of Kane's theory is its understanding of freedom as authorship rather than mere choice. Free will is not simply the capacity to select among alternatives but the capacity to become the ultimate source of one's moral identity. Moral responsibility therefore extends beyond isolated decisions to the historical process through which individuals gradually shape their character by making fundamental choices (Kane, 2005).

This historical conception of responsibility also explains why Kane rejects purely causal accounts of agency. Although actions may be causally attributable to particular individuals, such attribution is insufficient unless those individuals have also participated in forming the dispositions and commitments that motivate their actions. For this reason, Kane grounds moral responsibility in self-governing agency, according to which persons are not passive products of

external causal forces but active participants in the formation of their own moral identity (Kane, 1996; Clarke, 2003).

Kane further argues that indeterminism acquires philosophical significance only when it contributes to this process of self-formation. If indeterminism merely produces random outcomes, it weakens rather than supports moral responsibility. Its importance therefore lies in creating the conditions under which individuals can genuinely participate in shaping their future character through decisive choices (Kane, 1996, 126–145).

Kane's project is thus not simply a theory of free will but an account of the normative foundations of moral responsibility. His central question is not whether human beings possess metaphysical freedom in the abstract, but under what conditions they genuinely deserve praise or blame. He answers this question through three interrelated concepts, Ultimate Responsibility, Self-Forming Actions, and control, which together provide the theoretical framework for evaluating moral responsibility in both traditional contexts and emerging AI-mediated forms of decision-making.

Ultimate Responsibility

Ultimate Responsibility (UR) is the central concept in Robert Kane's theory of moral responsibility. Kane argues that individuals are fully morally responsible only if they are not merely the immediate causes of their actions but also the ultimate authors of the will and moral character from which those actions arise (Kane, 1996, 63–76). Moral responsibility therefore requires more than intentional action; it also requires that agents have actively participated in shaping the dispositions, values, and commitments that guide their conduct.

Kane develops UR in opposition to accounts that ground responsibility solely in the causal relation between agent and action. If an individual's beliefs, desires, and character are entirely determined by factors beyond personal control, attributing behaviour to that individual cannot justify moral praise or blame. Acting in accordance with one's desires is insufficient if one has played no meaningful role in their formation (Kane, 2005, 121–129). UR therefore distinguishes between merely performing an action and being its ultimate source. Responsibility extends beyond a particular decision to the historical process through which individuals freely shape their practical identity (Kane, 1996, 2005).

This historical dimension rests on two conditions. First, the action must be performed knowingly and freely. Second, the motivational structure underlying that action must ultimately derive from earlier choices in which the agent participated as a genuine author. If either condition is absent, the attribution of full moral responsibility becomes incomplete. UR is therefore not confined to isolated actions but encompasses the cumulative process through which moral character is formed over time (Kane, 2005).

Unlike psychological or behaviourist accounts, Kane treats moral character as something progressively constituted through successive free choices rather than as a fixed or predetermined condition. Indeterminism acquires normative significance only because it allows individuals to participate in forming the character for which they may later deserve praise or blame. In this respect, Ultimate Responsibility is fundamentally historical rather than momentary.

Despite its influence, UR has attracted substantial criticism. Fischer and Ravizza argue that moral responsibility requires only an appropriate form of guidance control rather than ultimate authorship of one's character (Fischer & Ravizza, 1998). Mele similarly questions whether indeterministic decisions can ground responsibility without introducing an unacceptable element of luck (Mele, 2006), while Pereboom argues that Ultimate Responsibility remains philosophically problematic even if determinism is rejected (Pereboom, 2014). More recently, Haji contends that indeterminism may weaken rather than secure agent control (Haji, 2022), and Campbell argues that Kane's later reformulation leaves unresolved how indeterministic efforts of will establish genuine authorship rather than contingent outcomes (Campbell, 2023).

These objections expose an important difficulty but do not, by themselves, undermine Kane's theory. Instead, they highlight the need to clarify the relationship between Ultimate Responsibility, indeterminism, and agent control. This issue becomes even more pressing in AI-mediated decision-making, where responsibility is exercised through technologically mediated forms of agency rather than through isolated individual choices.

Nevertheless, Ultimate Responsibility remains the normative core of Kane's account because it explains why individuals deserve praise or blame rather than merely describing the psychological conditions under which actions occur. Its central claim is that moral responsibility ultimately depends on whether persons can reasonably regard themselves as the authors of their own moral lives.

This conception also provides the point of departure for the present study. The question is not whether artificial intelligence replaces Ultimate Responsibility, but whether it changes the conditions under which that responsibility is exercised. As learning systems, predictive algorithms, and autonomous technologies become increasingly involved in practical deliberation, the relationship between authorship and responsibility becomes more complex. Yet complexity alone does not eliminate moral agency.

This conclusion is reinforced by Khazaei's ontological analysis of artificial agency. Although intelligent systems increasingly perform functions associated with agency, their functional autonomy is insufficient to establish genuine moral agency. A distinction must therefore be maintained between the operational agency of AI systems and the normative agency that grounds moral responsibility (Khazaei, 2026). AI may substantially reshape the conditions of

human deliberation without replacing human beings as the ultimate bearers of moral responsibility.

Accordingly, the criticisms directed at Kane's theory invite a reinterpretation rather than a rejection of Ultimate Responsibility. Instead of viewing it solely as a metaphysical condition governing isolated choices, it can be understood as a normative framework through which agents remain responsible for the gradual formation of their moral character. Read in this way, Kane's account provides a stronger basis for analysing moral responsibility in technologically mediated environments, where intelligent systems increasingly influence, but do not replace, human practical reasoning.

A useful illustration of this point can be found in the case of autonomous vehicles. Although real-time decisions are executed algorithmically, the ethically significant choices concerning the values embedded in these systems, such as how competing risks should be prioritized, are shaped by multiple human agents involved in their design, development, regulation, and deployment (Lin, Abney, & Jenkins, 2017). From a Kanean perspective, Ultimate Responsibility therefore remains attributable to these human agents rather than to the vehicle itself, because the system lacks the reflective and self-formative agency required for moral accountability (Kane, 1996, 2005). The example demonstrates that AI may possess operational autonomy while the deeper question of moral authorship remains connected to the human agents who shape the conditions, values, and commitments underlying technological decision-making.

Self-Forming Actions (SFAs)

The concept of Self-Forming Actions (SFAs) constitutes the principal mechanism through which Robert Kane explains Ultimate Responsibility. If Ultimate Responsibility specifies the condition under which individuals become the ultimate authors of their will and moral character, Self-Forming Actions explain how such authorship is achieved. According to Kane, moral character is neither fixed in advance nor simply the product of heredity and environmental influences. Rather, at certain critical moments individuals actively shape their future character through free choices. Kane refers to these formative decisions as Self-Forming Actions (Kane, 1996, 127–145)."

Kane distinguishes between ordinary decisions and those that fundamentally shape moral identity. Most everyday actions express an already established set of beliefs, values, and habits. By contrast, Self-Forming Actions occur when individuals confront genuine conflicts between competing values, where no alternative possesses decisive moral superiority. In resolving such conflicts, agents determine not only what they will do but also who they will become. The philosophical significance of SFAs therefore lies in their constitutive role in the formation of

moral character rather than in their immediate consequences (Kane, 1996; Kane, 2005, 135–145).

Within this framework, freedom is understood as the capacity for self-formation rather than merely the ability to choose between alternatives. Self-Forming Actions are therefore pivotal moments in the ongoing development of moral identity, and later actions derive much of their moral significance from these earlier formative choices. Moral responsibility is consequently inseparable from the historical development of the agent's character.

A distinctive feature of Kane's account is his interpretation of causal indeterminism. Rather than viewing indeterminism as a threat to agency, he argues that it is a necessary condition for genuine self-formation. If decisive choices were completely determined by prior causes, individuals could not genuinely be regarded as the authors of their own character. At the same time, Kane rejects the claim that indeterminism implies randomness. Decisions continue to arise from the agent's own beliefs, motivations, values, and deliberative efforts; indeterminism simply means that the future is not fully settled prior to the moment of choice (Kane, 1996, 2005).

This aspect of Kane's theory has generated substantial philosophical criticism. Mele (2006) argues that indeterminism risks weakening agency because, if more than one outcome remains genuinely possible under identical conditions, it becomes difficult to explain why agents deserve responsibility for one outcome rather than another. Similarly, Clarke (2003) contends that libertarian theories require a clearer explanation of how effective control is preserved under indeterministic conditions. Kane responds that indeterminism should be understood not as an external force acting upon the will but as an intrinsic feature of genuine deliberative conflict. Moral responsibility therefore depends on the agent's active participation in the process of self-formation rather than on the contingent occurrence of one particular outcome (Kane, 2021).

Although these criticisms identify important theoretical challenges, they do not diminish the explanatory significance of Self-Forming Actions. Instead, they highlight the need to clarify how self-formation and control interact in conditions of indeterminacy. For this reason, SFAs remain one of the most influential elements of Kane's account of moral responsibility.

For the purposes of the present study, the concept of Self-Forming Actions is particularly significant because contemporary AI systems increasingly participate in human deliberation through algorithmic recommendations, predictive analytics, and decision-support systems. This development raises an important philosophical question: can genuinely Self-Forming Actions still occur when intelligent systems contribute to decisions that shape an individual's future moral character? Rather than assuming that AI eliminates self-formation, the following analysis examines whether algorithmic mediation transforms the conditions under which Self-Forming Actions are exercised while preserving the human agent as the ultimate author of moral identity.

This question provides the point of departure for the subsequent evaluation of Self-Forming Actions in AI-mediated decision-making.

This issue can be illustrated through AI-assisted medical diagnosis. When a physician evaluates an AI recommendation, accepts it, rejects it, or modifies it in light of professional judgment and ethical considerations, the morally significant act is not simply the use of technological assistance but the agent's reflective engagement with that assistance. Such decisions may become part of the physician's moral development because they reveal how the agent determines the appropriate balance between reliance on external systems and independent professional judgment. From a Kanean perspective, AI-mediated choices can therefore constitute Self-Forming Actions rather than eliminate them, provided that the human agent remains actively involved in the deliberative process through which professional values and commitments are formed.

Kane's Later Revision: The "Fourth Way" and the Defense of Moral Agency

In *Making Sense of a Free Will that is Incompatible with Determinism: A Fourth Way Forward* (2021), Robert Kane offers his most recent defense of libertarian free will. Rather than proposing a new theory, he responds to longstanding criticisms concerning luck, agent control, and the intelligibility of incompatibilist freedom. Central to this response is what he calls the "Intelligibility Question": whether free will can be understood without reducing indeterminism to chance or metaphysical obscurity (Kane, 2021).

Kane argues that the traditional classification of libertarian theories into non-causal, agent-causal, and event-causal models fails to capture his position. His account combines elements of agent causation and causal processes while reducing neither to the other. Free actions arise through causal and psychological processes, yet these processes remain the agent's own, emerging from conflicts within the will. Moral agency is therefore neither external to the causal order nor reducible to it, but realized through self-formative deliberation within it (Kane, 2021).

The central aim of this reformulation is to show that indeterminism does not undermine control. Kane maintains that critics often equate indeterminism with randomness, whereas his view holds that indeterminism simply means the future is not completely fixed before choice. Self-Forming Actions remain expressions of the agent because they arise from deliberation among competing reasons and motivations. The "Fourth Way" is therefore best understood as a clarification of Ultimate Responsibility rather than a departure from Kane's earlier theory.

This clarification is particularly relevant to the present study. If moral responsibility can survive complex causal processes, then technological mediation does not by itself eliminate responsibility. What matters is that human agents continue to exercise reflective authority within those processes. This provides the conceptual basis for the later analysis of AI-assisted

decision-making, where Ultimate Responsibility depends on preserving the self-formative role of the human agent rather than excluding technological influence altogether.

Kane's revised account has nevertheless remained controversial. [Campbell \(2023\)](#) argues that it weakens the symmetry of competing efforts of will that characterized Kane's earlier theory and therefore fails to explain how indeterministic choices establish genuine authorship rather than contingent outcomes. [Haji \(2022\)](#) similarly maintains that introducing indeterminism at decisive moments continues to threaten, rather than secure, meaningful agent control.

These objections expose important philosophical tensions but do not undermine the central insight of Kane's theory. Instead, they suggest that Ultimate Responsibility is better understood as an ongoing process of self-formation than as a property of isolated decisions. This interpretation is especially valuable in AI-mediated environments, where responsibility increasingly depends on reflective decisions about the design, adoption, supervision, and evaluation of intelligent systems.

Viewed in this way, Kane's "Fourth Way" provides an important philosophical foundation for analysing moral responsibility in AI-mediated environments. Rather than undermining human agency, technological mediation reinforces the need to clarify how Ultimate Responsibility can be preserved under increasingly complex conditions of decision-making.

Evaluating Robert Kane's Theory of Moral Responsibility in the Age of Artificial Intelligence

Does Artificial Intelligence Undermine Ultimate Responsibility?

The principal challenge that artificial intelligence poses to Robert Kane's theory concerns not the autonomy of intelligent systems themselves but the status of Ultimate Responsibility in AI-mediated decision-making. Kane argues that moral responsibility is attributable only when individuals remain the ultimate source of their actions and moral character ([Kane, 1996; 2005](#)). As AI systems increasingly analyse information, generate recommendations, and influence practical deliberation, the question arises whether reliance on algorithmic assistance weakens this requirement.

At first sight, the answer appears affirmative. In fields such as medicine, finance, transportation, and defence, important decisions increasingly depend upon algorithmic recommendations. This development is reflected in Matthias's concept of the Responsibility Gap, according to which the complexity and unpredictability of learning systems make responsibility attribution increasingly difficult ([Matthias, 2004](#)).

Kane's account, however, suggests a different conclusion. Ultimate Responsibility does not require individuals to perform every stage of the decision-making process themselves. Rather,

it requires that they retain reflective authority over the acceptance, rejection, or revision of the reasons on which their decisions are based. External influences—including expert advice, institutional practices, and AI-generated recommendations—do not eliminate responsibility provided that agents remain responsible for their reflective judgments (Kane, 2005, 2021).

From this perspective, AI functions as an epistemic resource rather than a moral agent. It may generate information, predictions, or recommendations, but it does not possess the normative agency required for Ultimate Responsibility. This interpretation is consistent with Floridi and Sanders' distinction between functional and moral agency and with Coeckelbergh's claim that AI restructures rather than replaces human responsibility (Floridi & Sanders, 2004; Coeckelberg, 2020, 2023).

The increasing use of AI nevertheless requires a modest extension of Kane's framework. Responsibility begins not only with the final decision but also with earlier judgments concerning whether AI should be used, which systems are sufficiently reliable, and how their outputs ought to be evaluated. Ultimate Responsibility therefore extends across the governance of AI-assisted decision-making rather than being confined to the final act of choice.

From the perspective of the present study, artificial intelligence does not undermine Ultimate Responsibility but transforms the conditions under which it is exercised. Although AI reshapes the epistemic environment of decision-making, it lacks the capacity for self-forming agency that grounds moral responsibility in Kane's theory. The decisive philosophical issue is therefore not whether AI participates in decision-making, but whether human agents retain reflective authority over the adoption, evaluation, and supervision of algorithmic recommendations. As long as this authority remains with the human agent, Ultimate Responsibility continues to be attributed to the person rather than to the intelligent system.

Accordingly, the so-called Responsibility Gap is better understood as reflecting the increasing complexity of responsibility attribution in technologically mediated environments than as demonstrating the disappearance of human moral responsibility.

Self-Forming Actions and Algorithmic Mediation

Self-Forming Actions and Algorithmic Mediation Self-Forming Actions (SFAs), discussed in the previous chapter as the central mechanism of Kane's account of Ultimate Responsibility, provide the conceptual foundation for evaluating moral responsibility under AI-mediated conditions. Rather than reconstructing Kane's general theory, the present section examines whether Self-Forming Actions remain possible when intelligent systems increasingly participate in human deliberation and, if so, how their philosophical significance should be reinterpreted.

Within AI-mediated environments, Self-Forming Actions acquire renewed significance. Contemporary AI systems increasingly influence human deliberation by providing

recommendations, predictions, and probabilistic assessments. Yet these systems do not replace Self-Forming Actions; instead, they transform the conditions under which they occur. The morally significant question is no longer simply whether an algorithmic recommendation should be followed, but how individuals critically evaluate, interpret, and ultimately accept or reject it.

This reinterpretation becomes particularly clear in AI-assisted medical decision-making. Although advanced AI systems may analyse clinical data and provide diagnostic recommendations, physicians remain responsible for critically evaluating whether such recommendations should be accepted, modified, or rejected in light of the patient's circumstances and broader ethical considerations (Char et al., 2018; Topol, 2019). The morally significant element is therefore not the technological assistance itself, but the physician's reflective engagement with that assistance. Decisions concerning the appropriate degree of reliance on AI, the acceptance of algorithmic recommendations, and the preservation of independent professional judgment represent situations in which agents actively participate in shaping their practical identity. In this sense, AI-mediated medical decisions may constitute contemporary contexts for Self-Forming Actions because they involve reflective choices about values, responsibility, and professional commitments.

This account is consistent with contemporary technology ethics. Vallor argues that technologies shape the environments in which human virtues are cultivated, while Coeckelbergh maintains that artificial intelligence restructures rather than replaces human responsibility (Vallor, 2016; Coeckelbergh, 2020, 2023). Building on these insights, the present study argues that Self-Forming Actions in the age of artificial intelligence should not be understood only as conflicts between traditional moral values, but also as reflective decisions concerning how human agents engage with intelligent systems. AI therefore does not eliminate self-formation; rather, it creates new contexts in which human beings determine how technological assistance should be integrated into responsible judgment.

Artificial intelligence therefore does not eliminate the possibility of Self-Forming Actions; rather, it transforms the conditions under which they are exercised.

Kane's theory does not require that morally significant choices occur in the absence of external influences. Instead, it requires that agents retain the capacity to critically evaluate competing reasons and consciously endorse one course of action over another. In AI-assisted environments, algorithmic recommendations inevitably influence human deliberation, yet influence should not be confused with determination. A recommendation generated by an intelligent system does not become morally decisive unless it is reflectively accepted by the human agent. Consequently, Self-Forming Actions remain possible whenever individuals preserve the capacity for reflective endorsement or rejection of algorithmic outputs. In this

sense, artificial intelligence reshapes the context of self-formation without replacing the human agent as the author of moral character.

Agency, Control, and Moral Responsibility in Human-AI Interaction

Control constitutes a fundamental component of Robert Kane's theory of moral responsibility. Recent discussions likewise emphasize that moral responsibility depends on the appropriate exercise of control over action, even though philosophers continue to disagree about the precise nature of that control (Mele, 2021).

Although his framework is primarily developed through the concepts of Ultimate Responsibility and Self-Forming Actions, both concepts presuppose that agents exercise genuine control over their deliberative processes. Moral responsibility is therefore attributable only when individuals actively participate, through reflection and volition, in forming their decisions and making those decisions properly attributable to them (Kane, 1996; Kane, 2005).

Contemporary discussions of agency also emphasize that agency is not exhausted by the capacity to initiate actions but involves the exercise of practical judgment and the cultivation of dispositions that guide responsible action. From this perspective, virtues play a constitutive role in agency by shaping how agents deliberate, evaluate reasons, and respond to morally significant situations (Khazaei, 2019).

The integration of artificial intelligence complicates this requirement by introducing forms of decision-making in which individuals increasingly rely on algorithmic analyses. In many machine-learning systems, the reasoning processes underlying outputs remain partially opaque, creating the so-called black box problem. This raises an important question: can individuals retain sufficient control for moral responsibility when they cannot fully understand how algorithmic recommendations are generated?

From Kane's perspective, complete causal knowledge is not a necessary condition of responsibility. Control does not require agents to understand every mechanism involved in producing a decision; rather, it requires that they retain reflective authority over whether recommendations are accepted, modified, or rejected. Responsibility therefore depends not on generating all relevant information independently, but on critically evaluating its normative significance and incorporating it into one's own practical judgment (Kane, 2005, 121–129; Kane, 2021, 15–18).

This interpretation is consistent with contemporary accounts of AI ethics. Floridi and Sanders distinguish functional agency from moral agency, while Coeckelbergh argues that artificial intelligence restructures rather than replaces human responsibility (Floridi & Sanders, 2004; Coeckelbergh, 2020, 2023). Similarly, Levy maintains that moral responsibility depends upon an agent's conscious awareness of the morally relevant features of action rather than complete knowledge of every causal factor involved (Levy, 2014).

Human reliance on expert testimony provides an illuminating analogy. Individuals frequently make responsible decisions based on expert advice without fully understanding the underlying reasoning. AI recommendations operate in a comparable manner: the decisive issue is not algorithmic transparency alone, but whether human agents retain the capacity to evaluate, endorse, revise, or reject algorithmic outputs according to their own practical judgment.

Nevertheless, AI transforms the structure of control itself. Whereas Kane's original analysis focuses primarily on the agent's control at the moment of decision, AI-mediated environments require control throughout the entire process of technological engagement. Responsible agency now includes decisions about whether to employ AI systems, which systems to trust, how to assess their reliability, what degree of dependence is appropriate, and when human intervention remains necessary.

This broader understanding of control does not replace Kane's framework but extends its explanatory scope. The present study argues that control in AI-mediated environments should be understood not only as control over individual actions, but also as responsible governance of technological assistance. Even when algorithmic reasoning remains partially opaque, Ultimate Responsibility is preserved because responsibility extends across the process through which technological systems are selected, supervised, and integrated into human decision-making.

Accordingly, the central issue in AI-mediated moral responsibility is not whether artificial intelligence influences human decisions, but whether it undermines the conditions of free agency upon which Kane's theory depends. AI systems may provide information, predictions, and recommendations, but they do not possess the reflective capacity required for moral authorship. As long as human agents retain the ability to deliberate, endorse, reject, and revise algorithmic outputs, meaningful control and Ultimate Responsibility remain attributable to them.

The following section therefore examines whether contemporary AI systems genuinely threaten these conditions of free agency or whether they merely transform the environment in which human moral agency is exercised.

Evaluating Kane's Theory in the Age of AI

The preceding discussion indicates that Robert Kane's theory remains a philosophically compelling framework for understanding moral responsibility under conditions of AI-mediated decision-making. Although formulated before the emergence of contemporary artificial intelligence, its central concepts, Ultimate Responsibility, Self-Forming Actions, and reflective control, retain significant explanatory force when interpreted within technologically mediated environments (Kane, 1996, 2005, 2021). At the same time, the analysis suggests that these

concepts require conceptual refinement to address the distinctive features of human–AI interaction.

The present study argues that the decisive philosophical question is not whether artificial intelligence participates in human decision-making, but whether it undermines the conditions of free agency upon which Kane's account of moral responsibility depends. In Kane's framework, Ultimate Responsibility, Self-Forming Actions, and meaningful control presuppose that human agents retain the capacity for free and reflective choice. Consequently, AI would challenge Kane's theory only if it deprived human beings of their ability to deliberate independently, critically evaluate reasons, and exercise reflective control over their decisions.

Artificial intelligence does not invalidate Kane's account of Ultimate Responsibility. Algorithmic systems increasingly influence practical reasoning by generating recommendations, predictions, and probabilistic assessments, yet they do not become the authors of human action. Ultimate Responsibility depends on the agent's capacity to reflectively evaluate, endorse, or reject these outputs rather than on exclusive causal involvement in every stage of the decision-making process. Algorithmic assistance therefore modifies the epistemic conditions of deliberation without replacing human moral authorship.

A similar conclusion applies to Self-Forming Actions. Kane originally describes SFAs as decisions through which individuals shape their future moral character in situations of genuine value conflict. In AI-mediated environments, comparable conflicts arise when agents determine the appropriate degree of reliance on algorithmic advice, balance machine recommendations against independent judgment, or decide when automated outputs should be overridden. These decisions contribute directly to the formation of moral character and therefore represent contemporary contexts in which Self-Forming Actions continue to occur.

The notion of control likewise requires expansion. Kane correctly identifies reflective control as a necessary condition of moral responsibility, but AI extends control beyond the final moment of choice. Responsible agency also includes deciding whether AI should be employed, assessing the reliability of algorithmic systems, determining appropriate levels of dependence, and maintaining effective human oversight. Reflective control thus becomes a continuous process of technological engagement rather than a condition confined to isolated decisions.

This reinterpretation also reveals an important limitation of Kane's original framework. His account was primarily developed to explain the responsibility of individual agents, whereas AI-supported decisions frequently emerge within socio-technical systems involving developers, institutions, regulators, and users. Responsibility is therefore often distributed across multiple participants. Kane's theory alone cannot fully explain these institutional dimensions, although it continues to provide the normative conditions upon which distributed responsibility ultimately depends.

Contemporary AI ethics largely supports this interpretation. Floridi and Sanders distinguish functional agency from moral agency, Coeckelbergh argues that AI restructures rather than replaces human responsibility, and Vallor shows that technological environments influence the cultivation of moral character without themselves becoming moral agents (Floridi & Sanders, 2004; Coeckelbergh, 2020, 2023; Vallor, 2016). Together, these perspectives reinforce the conclusion that algorithmic systems may participate in practical reasoning while remaining incapable of bearing Ultimate Responsibility.

Under present technological conditions, there is no convincing philosophical reason to conclude that contemporary AI systems possess the capacity to eliminate human free agency. Although AI systems may significantly influence human deliberation, influence does not amount to compulsion. Human agents continue to retain the ability to reject, revise, or override algorithmic recommendations and therefore remain the ultimate bearers of moral responsibility in Kane's sense.

Rather than undermining Kane's libertarian account, artificial intelligence exposes the need to reinterpret its central concepts within technologically mediated contexts. Ultimate Responsibility, Self-Forming Actions, and reflective control remain the normative foundations of moral responsibility, but their exercise increasingly depends on how human agents govern, evaluate, and integrate algorithmic assistance into practical deliberation. Under these conditions, AI changes the context in which moral responsibility is exercised without displacing the human agent as its ultimate bearer.

Conclusion

The increasing integration of artificial intelligence into human decision-making has raised fundamental questions about whether classical theories of moral responsibility remain adequate for technologically mediated environments. This study has argued that Robert Kane's libertarian theory continues to provide a philosophically coherent normative framework for explaining human moral responsibility in the age of artificial intelligence, provided that its central concepts are reinterpreted in light of contemporary technological conditions.

The analysis demonstrates that the increasing role of artificial intelligence in human decision-making does not invalidate the core principles of Kane's theory. Rather than undermining Ultimate Responsibility, AI reshapes the practical conditions under which moral responsibility is exercised. Ultimate Responsibility, Self-Forming Actions, and reflective control remain philosophically applicable when interpreted within technologically mediated decision-making environments. Moral responsibility continues to depend upon the human agent's capacity to critically evaluate, supervise, endorse, or reject algorithmic recommendations and to remain the ultimate author of morally significant decisions.

At the same time, this study argues that Kane's concepts require conceptual extension to address AI-mediated decision-making. Ultimate Responsibility extends beyond the final decision to include earlier choices concerning whether and how AI should be used. Self-Forming Actions encompass reflective judgments about reliance on intelligent systems, while control extends beyond individual actions to include the responsible governance of technological assistance throughout the decision-making process.

The principal contribution of this study lies in demonstrating that Kane's theory and contemporary AI ethics should be understood as mutually reinforcing rather than competing approaches. By integrating libertarian theories of free will with current debates in AI ethics, this article proposes a conceptual reconstruction of libertarian moral responsibility for technologically mediated decision-making and demonstrates that artificial intelligence transforms the conditions under which responsibility is exercised without undermining its human foundation.

Nevertheless, Kane's framework also has important limitations. Because it was originally developed to explain individual moral agency, it cannot by itself fully explain responsibility within complex socio-technical systems involving developers, organizations, regulators, and users. Future research should therefore explore how Kane's account can be integrated with theories of distributed responsibility, collective agency, and AI governance.

The present analysis is limited to Kane's account of individual moral responsibility and does not attempt to provide a comprehensive theory of responsibility for all forms of artificial intelligence or all socio-technical systems. Moreover, if future forms of artificial intelligence were genuinely capable of eliminating the conditions necessary for human free agency, such as reflective deliberation, meaningful choice, and self-formation, the resulting challenge would extend beyond Kane's libertarian framework to many theories of moral responsibility that likewise presuppose human agency as a necessary foundation.

Overall, the emergence of artificial intelligence should not be understood as marking the end of human moral responsibility but as requiring a more sophisticated account of how responsibility is exercised within technologically mediated environments. When interpreted through this broader perspective, Kane's theory retains substantial philosophical and normative value, not because it eliminates the ethical challenges posed by AI, but because it provides a coherent framework for explaining why human beings remain the ultimate bearers of moral responsibility despite increasingly complex forms of technological assistance.

References

Campbell, N. (2024). Should Kane Abandon the Symmetry of Efforts of Will? *Journal of Philosophical Theological Research*, 26(3)5-24. [https://doi.org/ 10.22091/JPTR.2024.10366.3014](https://doi.org/10.22091/JPTR.2024.10366.3014)

- Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care Addressing Ethical Challenges. *New England Journal of Medicine*, 378(11), 981–983. <https://doi.org/10.1056/NEJMp1714229>
- Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford University Press.
- Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.
- Coeckelbergh, M. (2023). *The Political Philosophy of AI*. Polity Press.
- Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.
- Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379. <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>
- Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press.
- Gunkel, D. J. (2018). *Robot Rights*. MIT Press.
- Haji, I. (2022). Libertarianism and Luck. *Journal of Philosophical Theological Research*, (Special Issue on “Theology, Philosophy, and Ethics: In Memory of William J. Wainwright”), 24(3), 115-134. <https://doi.org/10.22091/jptr.2022.8297.2723>.
- Haji, I. (2023). Further Reflections on Lemos’s Indeterministic Weightings Model of Libertarian Free Action. *Journal of Philosophical Theological Research*, (Special Issue on "Comparative Philosophy of Religion"), 25(3), 119-130. <https://doi.org/10.22091/jptr.2023.9722.2929>
- Kane, R. (2021). *Making Sense of a Free Will that is Incompatible with Determinism: A Fourth Way Forward*. *Journal of Philosophical Theological Research*, 25, (3), 1–24. <https://doi.org/10.22091/JPTR.2021.7247.2595>
- Kane, R. (1996). *The Significance of Free Will*. Oxford University Press.
- Kane, R. (2005). *A Contemporary Introduction to Free Will*. Oxford University Press.
- Khazaei, Z. (2026). The Challenge of Agency in the Age of Artificial Intelligence: An Ontological Perspective. *Journal of Philosophical Theological Research*, 28(1), 7-14. <https://doi.org/10.22091/JPTR.2026.16299.3583>
- Khazaei, Z. (2019). Agency and virtues. *Journal of Philosophical Theological Research*, 21(3), 119–140. <https://doi.org/10.22091/jptr.2019.4673.2190>
- Kiener, M. (2025). AI and Responsibility: No Gap, but Abundance, *Journal of Applied Philosophy*, 42(1), 357-374, <https://doi.org/10.1111/japp.1276>
- Lemos, J. (2021). The Indeterministic Weightings Model of Libertarian Free Will. *Journal of Philosophical Theological Research*, (special issue on Free Will), 23(89), 137-156. <https://doi.org/10.22091/jptr.2021.7187.2582>
- Lemos, J. (2023). Haji and the Indeterministic Weightings Model of Libertarian Free Will. *Journal of Philosophical Theological Research*, (Special Issue on "Comparative Philosophy of Religion"), 25(3), 101-118. <https://doi.org/10.22091/jptr.2023.9648.2920>

- Levy, N. (2014). *Consciousness and Moral Responsibility*, Oxford University Press.
- Lin, P., Abney, K., & Jenkins, R. (Eds.). (2017). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford: Oxford University Press
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183. <https://doi.org/10.1007/s10676-004-3422-1>
- Mele, A. R. (2006). *Free Will and Luck*. Oxford University Press.
- Mele, A. R. (2021). Direct versus indirect: Control, moral responsibility, and free action. *Philosophy and Phenomenological Research*, 102(3), 559–573. <https://doi.org/10.1111/phpr.12680>
- Palmer, D. (2014). Free Will, Libertarianism, and Kane. In D. Palmer (Ed.), *Libertarian Free Will: Contemporary Debates* (pp. 2–11). Oxford University Press.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford University Press.
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25, 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.
- Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.