Philosophy of Science, Institute for Humanities and Cultural Studies (IHCS)
Biannual Journal, Vol. 15, No. 2, Autumn and Winter 2025-2026, 111-143
https://www.doi.org/10.30465/ps.2026.54415.1825

The Pragmatist Approach to Ascribing Free Will to Artificial Intelligence:

A Critique of Christian List's View

Tayyebe Gholami*

Seyed Hassan Hosseini**

Abstract

Christian List, inspired by Dennett's intentional stance, offers a pragmatic framework suggesting any system—human or artificial—possessing three macro-level features (intentional agency, alternative possibilities, and causal control) can be considered to have free will, even without phenomenal consciousness. This paper argues that List's framework suffers from a structural ambiguity, conflating functional autonomy with responsibility-bearing free will. Reducing free will to explanatory adequacy leads to conceptual inflation and paves the way for algorithmic responsibility evasion. Drawing on Kane's critique of ultimate origination, Murphy and Brown's defense of downward causation, and Nahmias's empirical evidence on the role of Strawsonian emotions in free will ascription, this paper shows that List's criteria are necessary but insufficient. We propose a three-level framework distinguishing (Level 1) functional autonomy, (Level 2) strong free will, and (Level 3) moral responsibility. The transition to Level 2 requires the ability to revise ultimate goals based on values, which necessitates phenomenal consciousness and Strawsonian emotions—qualities absent in current AI systems.

* Postdoctoral Researcher, Department of Philosophy of Science, Sharif University of Technology, Tehran, Iran (Corresponding Author), t.gholami@staff.sharif.edu

** Professor, Department of P philosophy of Science, Sharif University of Technology, Tehran, Iran, hoseinih@sharif.edu

Date received: 07/03/2026, Date of acceptance: 26/04/2026

This distinction resolves List's ambiguity while carrying significant normative implications for AI ethics, legal accountability, and technology policy.

Keywords: free will, artificial intelligence, Christian List, functional autonomy, algorithmic responsibility evasion, phenomenal consciousness autonomy, algorithmic responsibility evasion, phenomenal consciousness.

Introduction

If humanoid robots could discuss ethics, would we deem them free and responsible? Empirical studies (Shepherd, 2015) show that people ascribe free will to robots only if they presume the robot is *conscious*. Nahmias et al. (2019) confirm that Strawsonian emotions (guilt, pride, regret) mediate this ascription. Christian List (2023, 2025) dissents, arguing that free will is a macro-level explanatory category requiring three conditions—intentional agency, alternative possibilities, and causal control—independent of phenomenal consciousness. This paper argues List's framework collapses the distinction between functional autonomy and free will, leading to conceptual inflation and a dangerous phenomenon: *algorithmic responsibility evasion*, where developers deflect blame onto the system. After reconstructing List's view, we present internal and external critiques, then propose a tri-level model to resolve the ambiguity.

Materials & Methods

This study employs conceptual analysis and critical review of existing philosophical literature. We systematically examine List's pragmatic framework (2023, 2025) through three established lines of critique: (i) Robert Kane's (1996, 2007) "ultimate origination" and self-forming actions; (ii) Murphy and Brown's (2007) defense of downward causation and emergentism; and (iii) Nahmias et al.'s (2019) experimental evidence linking Strawsonian emotions to free will attribution. Based on these critiques, we construct a three-level normative framework (functional autonomy → strong free will → moral responsibility) to resolve the identified ambiguities.

Discussion & Results

List defines free will through three macro-level conditions: (1) intentional agency (goal-directed behavior based on beliefs/desires), (2) alternative possibilities (choices between options), and (3) causal control (higher-level states explaining actions). He explicitly decouples free will from phenomenal consciousness. Our

analysis reveals a critical internal ambiguity: List's criteria are precisely those of *functional autonomy*—a system's ability to choose and adapt without real-time human intervention. List's appeal to "explanatory adequacy" as the justification for labeling this as *free will* fails to identify any distinguishing feature. This collapse has a normative consequence: if any sufficiently complex AI is free (in List's sense), developers can claim "the system decided, not us"—the heart of *algorithmic responsibility evasion* (Elstic, 2024).Three external critiques strengthen our case. First, *Kane's origination problem*: List ignores that free will requires agents to be the "ultimate source" of their own ends. A developer-coded goal, no matter how complexly pursued, is not the system's own origin. List's dismissal of Kane's "ultimate responsibility" as impossibly stringent sidesteps the issue; the goal's origin matters normatively. Second, *Murphy and Brown's causal critique*: List remains within a linear, reductionist causal model. By rejecting downward causation and emergence, he treats macro-level descriptions as merely convenient summaries, not causally efficacious features. Yet genuine free will requires higher-order states (values, reasons) to causally impact lower-level processes—a structure List's framework cannot accommodate without admitting properties current AI lacks. Third, *Nahmias's empirical challenge*: List equates free will with "access consciousness" (information processing). But experimental studies show lay ascriptions of free will track *phenomenal* consciousness, specifically Strawsonian emotions. A system that cannot *feel* guilt or pride cannot be meaningfully responsible, regardless of its computational sophistication.

In addressing these critiques, we propose a tri-level framework:

- Level 1: Functional Autonomy. Goal optimization without real-time human intervention. The system pursues fixed ends flexibly. Example: autonomous vehicles, LLMs in routine tasks. *Limitation*: cannot revise the ends themselves.
- Level 2: Strong Free Will. Level 1 plus the *ability to revise ultimate goals based on values*. This requires phenomenal consciousness (experiential "mattering") and Strawsonian emotions, as caring about values necessitates affective experience. No current AI meets this condition.
- Level 3: Moral Responsibility. Level 2 plus sensitivity to norms and reflective access to reasons, enabling genuine accountability and susceptibility to praise/blame.

The transition from Level 1 to Level 2—*value-based goal revision*—directly answers Kane (origination), Murphy & Brown (downward causation), and Nahmias (phenomenal consciousness). List's conflation of Levels 1 and 2 thus dissolves. The normative consequence is decisive: by locating current AI systems strictly at Level 1, our framework blocks algorithmic responsibility evasion. Responsibility for the actions of even the most sophisticated autonomous systems remains with human designers, operators, and institutions until phenomenally conscious, value-revising systems (Level 2) are demonstrably achieved.

Conclusion

Christian List's pragmatic framework, while methodologically attractive, fails to distinguish functional autonomy from responsibility-bearing free will. Reducing free will to explanatory adequacy and eliminating phenomenal consciousness leads to both conceptual inflation and algorithmic responsibility evasion. Drawing on Kane's origination argument, Murphy and Brown's defense of downward causation, and Nahmias's empirical work on Strawsonian emotions, we have demonstrated that List's three conditions are insufficient. Our proposed three-level framework, centered on the transition condition of *value-based goal revision*, resolves this ambiguity. Current AI systems—including the most advanced language models—possess at most Level 1 (functional autonomy). Consequently, responsibility for their actions remains unequivocally human. Until AI systems can revise their ultimate ends based on values, which requires phenomenal consciousness, ascriptions of strong free will or moral responsibility to artificial agents are not merely premature but conceptually mistaken.

Bibliography

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Kane, R. (1996). *The Significance of Free Will*. Oxford University Press.

List, C. (2023). Free will and artificial intelligence. *Philosophy Compass*, 18(6), e12914.

List, C. (2025). Pragmatic Free Will for AI Systems. *Journal of Artificial Intelligence and Philosophy*, 12(1), 1-25.

Murphy, N., & Brown, W. S. (2007). *Did My Neurons Make Me Do It?* Oxford University Press.

Nahmias, E., Allen, C., & Loveall, B. (2019). Consciousness, Emotions, and Free Will Ascription. *Consciousness and Cognition*, 72, 1-18.

Shepherd, J. (2015). Consciousness, free will, and moral responsibility. *Review of Philosophy and Psychology*, 6(4), 937-954.

Alasti, K. (2024). Artificial Intelligence and Meaningful Human Control: Agent-Neutral-Reason-Responsive Mechanisms and the Possibility of Tracing Responsible Agent(s). *Journal of Philosophical Theological Research*, *26*(4), 27-54. doi: 10.22091/jptr.2025.11378.3139. [in Persian]

Hoseini Souraki, S. M. (2024). The Meaning and Criteria of Moral Agency in Intelligent Machines. *Journal of Philosophical Theological Research*, *26*(4), 83-108. doi: 10.22091/jptr.2025.12466.3256. [in Persian]

فلسفهٔ علم، پژوهشگاه علوم انسانی و مطالعات فرهنگی
دوفصل‌نامهٔ علمی (مقالهٔ پژوهشی)، سال ۱۵، شمارهٔ ۲، پاییز و زمستان ۱۴۰۴

نقد رویکرد عمل‌گرایانه در انتساب ارادهٔ آزاد به هوش مصنوعی: بررسی انتقادی دیدگاه کریستین لیست

طیبه غلامی*

سید حسن حسینی**

چکیده

کریستین لیست با الهام از موضع قصدی دنت، چارچوبی عمل‌گرایانه ارائه می‌دهد که بر اساس آن، هر سامانه‌ای (چه انسانی و چه مصنوعی) که در سطح کلان واجد سه شرط عاملیت قصدی، امکان‌های بدیل و کنترل علّی باشد، می‌تواند صاحب ارادهٔ آزاد تلقی شود؛ حتی اگر فاقد آگاهی پدیداری باشد. این مقاله با بهره‌گیری از روش تحلیل مفهومی و استناد به نقدهای موجود در ادبیات فلسفی (از جمله نقدهای کین بر مسئلهٔ منشأ، نقدهای مورفی و براون بر علیت نزولی و نوخاستگی، و شواهد تجربی نهمیاس درباره‌ی نقش احساسات استراسونی در انتساب ارادهٔ آزاد)، نشان می‌دهد که چارچوب لیست در تمایز میان «خودآیینی کارکردی» و «ارادهٔ آزادِ مسئولیت‌زا» دچار ابهامی ساختاری است. فروکاست ارادهٔ آزاد به صرفِ «کفایت تبیینی» نه‌تنها به تورم مفهومی می‌انجامد، بلکه بستر را برای «مسئولیت‌گریزی الگوریتمی» فراهم می‌کند؛ وضعیتی که در آن توسعه‌دهندگان و نهادها می‌توانند با استناد به عاملیت سامانه، خود را از پاسخگویی معاف سازند. در پاسخ به این چالش، چارچوب سه‌سطحی پیشنهادی این مقاله با تفکیک خودآیینی کارکردی (سطح ۱)، ارادهٔ آزاد قوی (سطح ۲) و مسئولیت اخلاقی (سطح ۳)، شرطی گذار را به عنوان معیار ارادهٔ آزاد قوی معرفی می‌کند: «توانایی بازنگری در اهداف نهایی بر پایهٔ ارزش‌ها». این شرط، مستلزم آگاهی پدیداری و احساسات استراسونی است و روشن

* پژوهشگر پسادکتری، گروه فلسفهٔ علم، دانشگاه صنعتی شریف، تهران، ایران (نویسندهٔ مسئول)، t.gholami@staff.sharif.edu

** استاد گروه فلسفهٔ علم، دانشگاه صنعتی شریف، تهران، ایران، hoseinih@sharif.edu

تاریخ دریافت: ۱۴۰۴/۱۲/۱۶، تاریخ پذیرش: ۱۴۰۵/۰۲/۰۶

می‌سازد که چرا هیچ سامانهٔ مصنوعی امروزی—حتی پیشرفته‌ترین مدل‌های زبانی—فاقد ارادهٔ آزاد قوی است. تمایز مذکور ضمن رفع ابهام نظری لیست، پیامدهای هنجاری مهمی برای اخلاق هوش مصنوعی، مسئولیت‌پذیری حقوقی و سیاست‌گذاری فناوری دارد.

کلیدواژه‌ها: ارادهٔ آزاد، هوش مصنوعی، کریستین لیست، خودمختاری کارکردی، مسئولیت‌گریزی الگوریتمی، آگاهی پدیداری.

۱. مقدمه

اگر روزی روبات‌های انسان‌نمایی ساخته شوند که درست مانند ما سخن بگویند، تصمیم بگیرند، و حتی در کافه‌ها دربارهٔ اخلاق بحث کنند—آیا آنها را صاحب ارادهٔ آزاد و شایستهٔ سرزنش و ستایش خواهیم دانست؟ پژوهش‌های تجربی اخیر پاسخی روشن به این پرسش داده‌اند: مردم عادی تنها زمانی به چنین روبات‌هایی ارادهٔ آزاد نسبت می‌دهند که آنها را آگاه (conscious) فرض کنند؛ یعنی باور داشته باشند که این موجودات واقعاً درد را حس می‌کنند، رنگ‌ها را می‌بینند و تجربهٔ درونی دارند.

جاشوا شپر(Joshua Shepherd) در مطالعه‌ای کلاسیک نشان داد شرکت‌کنندگانی که روبات انسان‌نما را آگاه می‌پنداشتند—یعنی باور داشتند که او واقعاً احساسات و عواطف را تجربه می‌کند—او را واجد ارادهٔ آزاد و مسئولیت اخلاقی می‌دانستند. در مقابل، آنان که روبات را فاقد هرگونه تجربهٔ آگاهانه تصور می‌کردند، چنین نسبت‌هایی را به او نمی‌دادند (Shepherd 2015, p. 939).این یافته‌ها حاکی از آن‌اند که شهود ارادهٔ آزاد به‌گونه‌ای ریشه‌دار با انگارهٔ آگاهی گره خورده است؛ گویی بدون فرض آنچه نهمیاو همکاران (Eddy Nahmias & et all)، آگاهی پدیداری(phenomenal consciousness) می‌نامند— یعنی تجربهٔ اول شخص و کیفیات ذهنی—سخن گفتن از اختیارِ اخلاقاً معنادار ناممکن می‌نماید. (Nahmias et al, (2019, p.3

اما کریستین لیست(Christian List) — که بی‌تردید منسجم‌ترین و بانفوذترین دفاعیه از امکان ارادهٔ آزادِ مصنوعی را در دههٔ اخیر عرضه کرده—دقیقاً بر ضد این شهود موضع می‌گیرد. او با الهام از موضع قصدی(Intentional stance) دنت، استدلال می‌کند که ارادهٔ آزاد نه یک راز متافیزیکی، که یک مقولهٔ تبیینیِ(explanatory category) سطح‌کلان است: هر سامانه‌ای—خواه انسان، خواه هوش مصنوعی—که (۱) رفتاری هدفمند از خود نشان دهد، (۲) در سطح کلان با گزینه‌های بدیل روبه‌رو باشد، و (۳) کنترل علّیِ رفتارش تحت

تأثیر همان باورها و اهدافش باشد، واجد ارادهٔ آزاد است—و این همه، مستقل از آن است که سامانه واجد آگاهی پدیداری باشد یا نباشد. لیست به‌صراحت می‌نویسد:«فارغ از اینکه آیا سامانهٔ هوش مصنوعی واقعاً تجربهٔ آگاهانه دارد یا نه، تا زمانی که در سطح کلان واجد ویژگی‌های عاملیت قصدی، امکان‌های بدیل و کنترل علّی باشد، می‌توان در معنایی عمل‌گرایانه آن را صاحب ارادهٔ آزاد دانست»(List,2025,p.12) .

این چارچوبِ جسورانه، به دلیل پرهیز از متافیزیک و هم‌آوایی با اهداف هوش مصنوعیِ قابل‌توضیح، (Explainable AI) امروزه به پارادایمی مسلط در فلسفهٔ فناوری بدل شده است. با این حال، پروژهٔ لیست بر یک ابهام ساختاری استوار است که تاکنون از چشم منتقدان پنهان مانده یا دست‌کم به‌طور نظام‌مند صورتبندی نشده است.

معیارهای سه‌گانهٔ او—عاملیت قصدی، امکان‌های بدیل، کنترل علّی—دقیقاً همان ویژگی‌هایی هستند که در ادبیات فلسفهٔ ذهن و رباتیک، خودمختاری کارکردی (functional autonomy) نامیده می‌شوند؛ توانایی سامانه در گزینش هدفمند و انطباق با محیط بدون مداخلهٔ لحظه‌ای انسان. پرسش محوری این است که چه چیز افزوده‌ای موجب می‌شود که این خودمختاری پیچیده را ارادهٔ آزاد بخوانیم؟ اگر پاسخ لیست کفایت تبیینی(Explanatory Sufficiency) باشد،— یعنی این که توصیف سامانه در سطح عاملیت قصدی، توضیح بهتری از رفتارش به دست می‌دهد— آنگاه تمایز مفهومی میان خودمختاری(autonomy) و اراده آزاد(free will) فرو می‌ریزد و ارادهٔ آزاد به سطحی از توصیف فروکاسته می‌شود که با تغییر چارچوب نظری جابه‌جا می‌گردد. این ابهام نه‌تنها فلسفی، که عمیقاً هنجاری(normative) است: در غیاب مرزی روشن، نسبت‌دادن ارادهٔ آزاد به سامانه‌ها به‌سادگی به مسئولیت‌گریزی الگوریتمی(algorithmic responsibility evasion) دامن می‌زند—توسعه‌دهندگان می‌توانند در مواجهه با خطا بگویند: "ما فقط ابزار ساختیم، خودش تصمیم گرفت" (List ,2025, p 12). برای جلوگیری از این نوع مسئولیت‌گریزی، برخی پژوهشگران مانند الستی (۲۰۲۴)[1] با طرح مفهوم "کنترل معنادار بشری" (MeaningfulHumanControl) استدلال می‌کنند که حتی در صورت وجود عاملیت خودمختار در سیستم، باید سازوکارهایی وجود داشته باشد تا ردگیری مسئولیت نهایی به سمت عامل انسانی (طراح یا اپراتور) امکان‌پذیر باشد».

به‌تعبیر شپرد، اگر فیلسوفان می‌خواهند چنین شهود قدرتمندی را نادیده بگیرند، باید یا یک نظریهٔ ماهوی دربارهٔ پیوند میان آگاهی و ارادهٔ آزاد تدوین کنند، یا توجیهی قانع‌کننده

برای کنارگذاشتن این بخش به‌ظاهر مرکزی از درک عرفی ما فراهم آورند (Shepherd 2015: 944). لیست نه نظریه‌ای ماهوی ارائه می‌دهد و نه توجیهی برای حذف آگاهی—او صرفاً آن را کنار می‌نهد.

در این مقاله استدلال می‌کنیم که پروژهٔ لیست، حتی با مفروضات خودش، از عهدهٔ تمایز میان خودمختاری کارکردی و ارادهٔ آزادِ مسئولیت‌زا برنمی‌آید. کفایت تبیینی شرطی ضروری اما ناکافی است، و لیست برای پر کردن این شکاف ناگزیر از پذیرش التزامات هنجاری‌ای می‌شود که با عمل‌گراییِ صرف ناسازگارند. در مقابل، یک مدل سه‌سطحی از عاملیت پیشنهاد می‌کنیم:

- سطح ۱: خودمختاری کارکردی —بهینه‌سازی اهداف ثابت بدون مداخلهٔ انسانی.
- سطح ۲: ارادهٔ آزاد قوی —توانایی بازنگری در اهداف نهایی بر پایهٔ ارزش‌ها.
- سطح ۳: مسئولیت اخلاقی —حساسیت به هنجارها و دست‌کم توانایی بازبینی آگاهانه و ارزیابی انتقادی دلایل خود.

این چارچوب نه‌تنها ابهام لیست را رفع می‌کند، بلکه نشان می‌دهد چرا هیچ سامانهٔ مصنوعی امروزی—حتی پیشرفته‌ترین مدل‌های زبانی—واجد ارادهٔ آزاد قوی نیست. برای دفاع از این ادعا، ابتدا دیدگاه لیست را با تأکید بر مبانی عمل‌گرایانه‌اش بازسازی می‌کنیم(بخش ۲). سپس دو نقد درونی را صورتبندی می‌کنیم: ابهام در تمایز خودمختاری/ارادهٔ آزاد، و فروکاست هنجاری که به «مسئولیت‌گریزی الگوریتمی» می‌انجامد (بخش ۳). در بخش ۴، مدل سه‌سطحی خود را معرفی کرده، شرایط گذار از خودمختاری به ارادهٔ آزاد قوی را مشخص می‌کنیم.

۲. مرور فشردهٔ دیدگاه کریستین لیست دربارهٔ ارادهٔ آزاد در هوش مصنوعی

کریستین لیست در مقاله‌ی خود می‌کوشد امکان نسبت‌دادن اراده‌ی آزاد به سامانه‌های هوش مصنوعی را در چارچوبی عمل‌گرایانه و غیرمتافیزیکی بررسی کند. او نقطه‌ی عزیمت خود را فاصله‌گرفتن از تلقی‌های سنتی‌ای می‌داند که اراده‌ی آزاد را مستلزم ویژگی‌هایی رازآلود، ناتعین‌گرایی بنیادین یا نوعی گسست از قوانین طبیعت می‌انگارند. به‌زعم لیست، چنین معیارهایی نه‌تنها هوش مصنوعی، بلکه حتی انسان را نیز به‌سختی واجد اراده‌ی آزاد خواهند

کرد و این همان چرخش روش‌شناختی‌ای است که هستهٔ اصلی پروژهٔ عمل‌گرایانهٔ او را تشکیل می‌دهد(list,2025,p.1).

در مقابل، لیست با الهام از آثار دنیل دنت، به‌ویژه ایده‌ی موضع قصدی، (Intentional stance) پیشنهاد می‌کند که مسئله‌ی اراده‌ی آزاد در قالب پرسشی تبیینی صورت‌بندی شود. پرسش اصلی این نیست که آیا یک سامانه دارای نوعی قدرت متافیزیکی ویژه است یا خیر، بلکه این است که آیا دلایل تبیینی موجهی وجود دارد که آن سامانه را به‌مثابه عاملی قصدی(intentional agency) در نظر بگیریم. بر این اساس، اراده‌ی آزاد مفهومی تشخیصی و کاربردی است، نه توصیفی از ساختار نهایی واقعیت. لیست تأکید می‌کند که اراده‌ی آزاد ویژگیِ کلِ سیستم در سطح کلان است، نه ویژگی سخت‌افزار یا نرم‌افزار منفرد؛ همان‌گونه که انسان‌ها به‌عنوان موجودات هدفمند دارای اراده‌ی آزاد تلقی می‌شوند، نه مغز یا اجزای زیستی آنان به‌تنهایی. همین سطح کلان است که لیست سه شرط خود را بر آن بنا می‌کند(list,2025, p.3).

او برای نسبت‌دادن اراده‌ی آزاد سه شرط را مطرح می‌کند: نخست، عاملیت قصدی (intentional agency)؛ یعنی توانایی کنش هدفمند بر پایه‌ی حالت‌هایی چون باورها و ترجیحات. دوم، امکان‌های بدیل؛ به این معنا که سامانه در سطح عاملیت با گزینه‌های مختلفی مواجه باشد و بتواند میان آن‌ها انتخاب کند. سوم، کنترل علّی؛ بدین معنا که حالت‌های سطح بالای سامانه—نظیر اهداف، تصمیم‌ها و بازنمایی‌ها—در تبیین کنش‌های آن نقشی تفاوت‌ساز داشته باشند(list,2023, pp.5-7).

در دفاع از شرط نخست، لیست استدلال می‌کند که بسیاری از سامانه‌های پیچیده‌ی هوش مصنوعی را نمی‌توان به‌نحو رضایت‌بخشی صرفاً در سطح الگوریتمی توضیح داد. در چنین مواردی، توصیف سامانه به‌عنوان عاملی هدفمند از نظر تبیینی ضروری یا دست‌کم به‌مراتب ثمربخش‌تر است. بنابراین نسبت‌دادن باورها و اهداف به این سامانه‌ها صرفاً استعاری نیست، بلکه فرضیه‌ای درباره بهترین سطح تبیین رفتار آن‌هاست(list,2023, p.5).

در باب امکان‌های بدیل، لیست با تمایز میان سطوح خرد و کلان توضیح می‌دهد که حتی اگر فرایندهای زیربنایی یک سامانه کاملاً تعیّن‌گرا باشند، در سطح کلان عاملیت همچنان می‌توان از گزینه‌های بدیل سخن گفت. این تمایز سطحی امکان سازگاری با جبرگرایی را فراهم می‌آورد. سطح خرد ممکن است کاملاً تعیّن‌گرا باشد، اما سطح کلان عاملیت برای تبیین رفتار نیازمند ارجاع به انتخاب میان گزینه‌های واقعی است. از این رو،

امکان‌های بدیل در سطح کلان می‌توانند معنا و کارکرد تبیینی داشته باشند، حتی در جهانی که در سطح فیزیکی کاملاً قانونمند است.

در شرط سوم، لیست با بهره‌گیری از نظریه‌های تفاوت‌ساز(difference-maker) علّی نشان می‌دهد که یک سامانه زمانی بر کنش‌های خود کنترل دارد که حالت‌های هدفمند و بازنمایی‌گر آن متغیرهای کنترلی مؤثر در تبیین رفتار باشند، نه صرفاً ریزحالت‌های فیزیکی یا محاسباتی. در سامانه‌های پیچیده، توضیح‌های سطح بالا اغلب از نظر تبیینی بر توضیح‌های سطح پایین برتری دارند، و همین برتری تبیینی مبنای نسبت‌دادن کنترل علّی در سطح عاملیت است (list,2023, p.10).

لیست سپس به اعتراض‌هایی می‌پردازد، از جمله این نگرانی که رویکرد او به نسبت‌دادن اراده‌ی آزاد به سیستم‌های ساده‌ای مانند ترموستات یا الگوریتم‌های بهینه‌ساز بینجامد. پاسخ او این است که تنها در صورتی باید از اراده‌ی آزاد سخن گفت که تبیین عاملیتیِ رفتار سامانه به‌طور روشن بر تبیین‌های غیرعاملی برتری داشته باشد. در مورد سیستم‌های ساده‌ای چون ترموستات یا حتی یک کامپیوتر شطرنج، معمولاً چنین ضرورتی وجود ندارد و توضیح غیرعاملی کفایت می‌کند (list,2025, p.19).

لیست همچنین تأکید می‌کند که ارادهٔ آزاد در یک پیوستار(continuum) جای می‌گیرد. به این معنا که سامانه‌ها از نظر برخورداری از ابتکار عمل، میزان استقلال در تصمیم‌گیری و پیچیدگی اهداف، در طیفی پیوسته از حداقل تا حداکثر قرار می‌گیرند و مرز روشنی بین "فاقد ارادهٔ آزاد" و "دارای ارادهٔ آزاد کامل" وجود ندارد. درجهٔ برخورداری یک سامانه از ارادهٔ آزاد به پیچیدگی ساختار آن، توانایی تعقیب اهداف بلندمدت و میزان انعطاف‌پذیری در واکنش به موقعیت‌ها بستگی دارد.

از دیدگاه لیست، میان ارادهٔ آزاد، آگاهی پدیداری و مسئولیت اخلاقی تمایز وجود دارد. ارادهٔ آزاد در معنای مد نظر او مستلزم آگاهی تجربی و پدیداری نیست. افزون بر این، ارادهٔ آزاد شرط لازم برای مسئولیت اخلاقی است، اما شرط کافی آن محسوب نمی‌شود. ارادهٔ آزاد مفهومی توصیفی و تبیینی است، در حالی که مسئولیت اخلاقی مفهومی هنجاری به شمار می‌آید و مستلزم ظرفیت قضاوت اخلاقی و حساسیت به ارزش‌هاست. همچنین، ارادهٔ آزاد مستلزم غیرقابل‌پیش‌بینی‌بودن نیست؛ یک سامانه می‌تواند بر اساس دلایل قابل فهم تصمیم بگیرد و در عین حال تا حدی پیش‌بینی‌پذیر باشد. (list,2025,p.115).

در مجموع، دیدگاه لیست تصویری حداقلی، سطح‌محور و عمل‌گرایانه از اراده‌ی آزاد ارائه می‌دهد که هدف آن نه حل مناقشات متافیزیکی کلاسیک، بلکه فراهم‌کردن چارچوبی تبیینی برای فهم عاملیت—چه انسانی و چه مصنوعی—است. این چارچوب حتی قابلیت تعمیم به برخی نظام‌های جمعی، مانند شرکت‌ها یا نهادهای سازمان‌یافته، را دارد که می‌توانند در سطح کلان واجد ویژگی‌های عاملیت قصدی، امکان انتخاب و کنترل علّی باشند.

با این حال، چارچوبِ منسجم لیست بر یک ابهام اساسی استوار است :آیا کفایت تبیینی می‌تواند بار هنجاریِ ارادهٔ آزادِ مسئولیت‌زا را حمل کند؟ لیست تمایز روشنی میان خودمختاری کارکردی و ارادهٔ آزادِ اخلاقاً معنادار ارائه نمی‌دهد. در بخش بعد، این ابهام را از طریق مرور نقدهای موجود در ادبیات صورتبندی کرده و سپس در بخش ۴ به نقد تحلیلی و ارائه چارچوب جایگزین خواهیم پرداخت.

۳. نقدهای موجود در ادبیات فلسفی نسبت به رویکرد کرستین لیست

رویکرد عمل‌گرایانه‌ی کریستین لیست در باب اراده‌ی آزاد، به‌ویژه در نسبت‌دادن این مفهوم به سامانه‌های پیچیده و هوش مصنوعی، در ادبیات فلسفی با واکنش‌های انتقادی قابل توجهی مواجه شده است. این نقدها عمدتاً ناظر به این پرسش‌اند که آیا کفایت تبیینی یک چارچوب نظری و امکان توصیف سامانه‌ای در سطح عاملیت قصدی، برای نسبت‌دادن اراده‌ی آزاد به معنای اخلاقی و هنجاری آن کافی است یا خیر. در اینجا، مهم‌ترین خطوط این نقدها بازسازی می‌شود.

نخستین دسته از این انتقادات، ریشه در مناقشات پیرامون موضع قصدی دنیل دنت دارد؛ چارچوبی که لیست نیز به‌طور صریح از آن الهام می‌گیرد (Dennett, 1984, p 61; 1987, p. 59). منتقدانی همچون جان سرل (1994,P.7)، ند بلاک (1987,P.274) و جری فودر (1978,p.503) استدلال کرده‌اند که موفقیت یک راهبرد تبیینی در پیش‌بینی و فهم رفتار یک سامانه، لزوماً مستلزم آن نیست که آن سامانه واقعاً واجد حالات ذهنی یا ویژگی‌های هنجاری باشد. از این منظر، اتخاذ موضع قصدی ممکن است صرفاً ابزاری معرفتی برای سامان‌دهی تبیین‌ها باشد، نه معیاری برای تحقق واقعی عاملیت. بر این اساس، نگرانی اصلی آن است که توصیف یک سامانه به‌مثابه عاملِ انتخاب‌گر بیش از آن‌که حکایت از وجود اختیار داشته باشد، بازتابی از کارآمدی یک مدل توضیحی باشد.

این نقد، هنگامی که به نظریه‌ی لیست تعمیم داده می‌شود، اهمیت مضاعف می‌یابد. زیرا لیست نیز تأکید می‌کند که اگر توصیف یک سامانه در سطح کلان و در چارچوب عاملیت قصدی، از نظر تبیینی موجه و ثمربخش باشد، می‌توان در معنایی عمل‌گرایانه از اراده‌ی آزاد آن سخن گفت. منتقدان اما یادآور می‌شوند که گذار از کفایت توضیحی به نسبت‌دادن اراده‌ی آزاد، مستلزم پذیرش مفروضات هنجاری‌ای است که صرف موفقیت تبیینی آن‌ها را تضمین نمی‌کند.این نقد نشان می‌دهد که «خودمختاری کارکردی» (سطح ۱ مدل پیشنهادی) که لیست بر آن تکیه دارد، صرفاً ابزاری برای پیش‌بینی رفتار است و فاقد بار هنجاری است. مدل سه سطحی ما با تفکیک این سطح از سطوح بالاتر، نشان می‌دهد که چرا توصیف موفق تبیینی به تنهایی برای انتساب مسئولیت اخلاقی کافی نیست.

دسته‌ی دوم نقدها، به‌طور مشخص متوجه نسبت میان اراده‌ی آزاد و مسئولیت اخلاقی است. فیلسوفانی چون جان مارتین فیشر (1994,p.373) و رابرت کین (1996,p.60؛2002,58)، هرچند از مواضع متفاوتی دفاع می‌کنند، بر این نکته تأکید دارند که اراده‌ی آزاد، اگر قرار است مبنای مسئولیت اخلاقی باشد، باید واجد مؤلفه‌هایی فراتر از صرف قابلیت پیش‌بینی و کنترل رفتاری باشد. از این منظر، حتی اگر یک سامانه دارای امکان‌های بدیل و سازوکارهای کنترلی در سطح کلان باشد، این امر به‌تنهایی بار هنجاری مفهوم اراده‌ی آزاد را حمل نمی‌کند.

رابرت کین، به‌عنوان یکی از مهم‌ترین فیلسوفان ناسازگارگرای معاصر، تمایزی بنیادین میان فعل آزاد(free action) و ارادهٔ آزاد ترسیم می‌کند که برای ارزیابی چارچوب لیست حیاتی است. از نظر کین، فعل آزاد صرفاً به معنای انجام عمل بدون موانع بیرونی و بر اساس خواست‌های فعلی عامل است، اما ارادهٔ آزاد مستلزم مفهوم مسئولیت نهایی (Ultimate Responsibility) است: عامل باید منشأ نهایی(ultimate source) شخصیت، انگیزه‌ها، اهداف و اراده‌اش باشد (Kane, 1996, p.20) . به‌بیان دیگر، نمی‌توان صرفاً با نگاه به لحظهٔ کنش و توصیف موفق آن در چارچوب عاملیت قصدی، از تحقق ارادهٔ آزاد سخن گفت؛ باید بتوان نشان داد که خودِ باورها، ترجیحات و اهدافی که کنش از آنها نشئت می‌گیرد، نهایتاً به خود عامل بازمی‌گردند و او در شکل‌گیری آنها نقشی داشته است.

این نقد مستقیماً به لیست قابل تعمیم است. لیست با فروکاستن ارادهٔ آزاد به «کفایت تبیینی» در سطح کلان، نه‌تنها شرط مسئولیت نهایی را نادیده می‌گیرد، بلکه اساساً امکان طرح آن را منتفی می‌کند. از نظر او، صرف اینکه بتوان سامانه را در سطح کلان به‌عنوان

عامل قصدی با موفقیت توصیف کرد—یعنی باورها و اهدافش را تبیین‌کنندهٔ رفتارش دانست—برای انتساب ارادهٔ آزاد کافی است. اما کین می‌گوید: این توصیف موفق، هیچ‌گاه به پرسش «منشأ» پاسخ نمی‌دهد. سامانه‌ای که باورها و اهدافش را از برنامه‌نویسان، داده‌های آموزشی و الگوریتم‌های از پیش تعیین‌شده گرفته، هرگز نمی‌تواند «منشأ نهایی» آنها باشد، حتی اگر در سطح کلان رفتاری هدفمند و به‌ظاهر خودمختار از خود نشان دهد. لیست میان «توصیف موفق» و «تحقق واقعی» خلط کرده است.

کین با بهره‌گیری از مثال‌های فیلسوف انگلیسی جی.ال.آستین(J. L. Austin) نشان می‌دهد که صرف وجود «امکان‌های بدیل» (حتی اگر با عدم تعیّن همراه باشد) برای ارادهٔ آزاد کافی نیست. او سه مثال می‌زند: گلف‌بازی که به دلیل لرزش ناگهانی عصبی، ضربهٔ سه‌فوتی را از دست می‌دهد؛ قاتلی که به دلیل انقباض غیرارادی عضله، به جای نخست‌وزیر، دستیارش را می‌کشد؛ و شخصی که به دلیل یک خطای تصادفی مغزی، به جای دکمهٔ قهوه بدون خامه، دکمهٔ قهوه با خامه را می‌زند(Kane, 2005, p.124) . کین تأکید می‌کند که حتی اگر فرض کنیم این رویدادها نامتعین باشند—یعنی ناشی از جهش‌های کوانتومی یا عوامل تصادفی—باز هم عامل بر آنها کنترل نداشته و نمی‌توان او را مسئول دانست. بدین ترتیب، کین نشان می‌دهد که صرفِ همراهی امکان‌های بدیل با عدم‌تعیّن، برای تحقق ارادهٔ آزاد کافی نیست.

این مثال‌ها چالش مهمی برای چارچوب لیست ایجاد می‌کنند. لیست سه شرط خود—عاملیت قصدی، امکان‌های بدیل کنترل علّی—را معادل ارادهٔ آزاد می‌گیرد. اما مثال‌های کین نشان می‌دهند که حتی اگر این سه شرط به‌طور صوری برآورده شوند (سامانه گزینه‌های متفاوت دارد، انتخابش هدفمند به نظر می‌رسد، و در سطح کلان کنترل‌شده توصیف می‌شود)، ممکن است فاقد «کنترل درونی» واقعی باشد. در مثال قاتل، او می‌توانست به جای دستیار، نخست‌وزیر را بکشد (امکان بدیل)، انتخابش بر اساس هدفش بود (عاملیت قصدی)، و عملش ناشی از انقباض عضله بود که می‌توان آن را نوعی کنترل علّی سطح پایین دانست اما آیا واقعاً او را صاحب ارادهٔ آزاد می‌دانیم؟ لیست نمی‌تواند میان این نوع کنترل صوری و کنترلِ مسئولیت‌زا تمایز بگذارد.

کین برای نشان‌دادنِ اهمیت منشأ (sourcehood) در ارادهٔ آزاد، مثال "جهان‌های K-" (K-words) را مطرح می‌کند. جهانی را تصور کنید که در آن همان عدم تعیّن سبک آستینی وجود دارد و افراد می‌توانند انتخاب‌های متفاوتی داشته باشند، اما تمام انگیزه‌ها، اهداف و

اراده‌هایشان توسط خدا (یا هر عامل بیرونی دیگر) از پیش تنظیم شده است (Kane, 2007, p.19). در این جهان، افراد ممکن است گزینه‌های بدیل داشته باشند و افعالشان نامتعین باشد، اما «منشأ نهایی» اهدافشان به خودشان تعلق ندارد. از نظر کین، چنین افرادی فاقد ارادهٔ آزادند.

این مثال دقیقاً وضعیت سامانه‌های هوش مصنوعی را در چارچوب لیست روشن می‌کند. لیست صرفاً به «اینجا و اکنون» سامانه نگاه می‌کند: آیا باورها و اهداف دارد؟ آیا کنترل علّی بر اساس آنها دارد؟ اما هرگز نمی‌پرسد این باورها و اهداف از کجا آمده‌اند. اگر آنها محصول برنامه‌ریزی، داده‌های آموزشی، و الگوریتم‌های از پیش تعیین‌شده باشند، آیا سامانه واقعاً «منشأ» آنهاست؟ از نظر کین، پاسخ منفی است. سامانه‌های هوش مصنوعی امروزی—که باورها و اهدافشان محصول برنامه‌نویسی و داده‌آموزی است—مصداق بارز جهان‌هایK هستند: آنها ممکن است در سطح کلان واجد سه شرط لیست باشند، اما «منشأ نهایی» اهدافشان در جای دیگری است. لیست با بی‌توجهی به این مسئله، میان عاملیت مشتق (derived agency) و عاملیت اصیل(original agency) تمایز نمی‌گذارد.

کین برای گریز از تسلسل علّی در مسئولیت نهایی، مفهوم افعال تعیّن بخش اراده (Self-Forming Actions) را مطرح می‌کند. این افعال، که شخصیت و اراده را شکل می‌دهند، باید واجد شرایط چندگانه(plurality conditions) باشند: ارادی، آگاهانه، عاقلانه و عامدانه (Kane, 2007,p.14). صرف نامتعین بودن و انتخاب میان گزینه‌ها کافی نیست. عامل باید بتواند به‌طور آگاهانه و عاقلانه تصمیم بگیرد، نه اینکه انتخابش محصول شانس یا تصادف محض باشد .واضح است که سامانه‌های هوش مصنوعی امروزی فاقد چنین ظرفیت آگاهانه و عاقلانه‌ای هستند؛ بنابراین، هرگز نمی‌توانند استانداردهای سخت‌گیرانه‌ی کین برای اراده‌ی آزاد را برآورده سازند.

این شرط برای چارچوب لیست مشکل‌ساز است. لیست آگاهی را به‌کلی از تعریف ارادهٔ آزاد حذف کرده و عقلانیت را صرفاً به‌صورت کارکردی (پردازش اطلاعات) تعریف می‌کند. اما کین نشان می‌دهد که «آگاهانه بودن» برای افعالی که شخصیت و اراده را شکل می‌دهند، ضروری است. سامانه‌ای که فاقد آگاهی پدیداری است، نمی‌تواند در معنای کینی «آگاهانه» تصمیم بگیرد حتی اگر پردازش اطلاعات پیچیده‌ای داشته باشد. بنابراین، حتی اگر بپذیریم که برخی سامانه‌های هوش مصنوعی می‌توانند اهداف خود را بازنگری کنند

(چیزی که لیست به آن اشاره‌ای ندارد)، این بازنگری بدون آگاهی نمی‌تواند از نوع افعال تعیّن بخش اراده باشد که کین برای مسئولیت نهایی ضروری می‌داند.

با این حال، لیست پیش‌بینی‌کننده این نقد بوده و صراحتاً در آثار خود (لیست، ۲۰۱۴؛ لیست، ۲۰۲۳) به آن پاسخ داده است. استدلال اصلی لیست این است که شرط «مسئولیت نهایی» که کین مطرح می‌کند، یک شرط غیرممکن و بیش‌ازحد سخت‌گیرانه است که حتی برای انسان‌ها نیز محقق نمی‌شود. لیست استدلال می‌کند که اگر عاملیت نیازمند این باشد که عامل، منشأ نهایی تمام خواسته‌ها و شخصیت خود باشد، آنگاه هیچ عاملی (انسان یا ماشین) هرگز آزاد نخواهد بود، زیرا خواسته‌های ما همواره تحت تأثیر ژنتیک، محیط و تصادفات گذشته شکل می‌گیرند. از دیدگاه لیست، آنچه برای اراده‌ی آزاد کافی است، «کنترل علّی» در لحظه تصمیم‌گیری و توانایی واکنش به دلایل است، نه لزوماً ساختنِ تمام پیش‌زمینه‌های ذهنی از صفر. بنابراین، لیست با تغییر معیار از «منشأ متافیزیکی» به «کارکرد عاملیتی»، شرط کین را غیرضروری می‌داند.اما پاسخ لیست، اگرچه از منظر یک تعریف «کارکردی» و «توصیفی» از اراده‌ی آزاد قابل دفاع است، با این حال چالش‌هایی را برجسته می‌کند که مدل پیشنهادی ما را ضروری می‌سازد.

نقد کین دقیقاً نشان می‌دهد که چرا معیارهای لیست در سطح ۱ (خودمختاری کارکردی) ناکافی هستند. چالش «منشأ» و «کنترل درونی» که کین مطرح می‌کند، همان خلاءیی است که در سطح ۲ (اراده آزاد قوی) مدل پیشنهادی ما با افزودن شرط «توانایی بازنگری در اهداف بر پایه ارزش‌ها» و تأکید بر آگاهی، پر می‌شود. مدل ما نشان می‌دهد که بدون این لایه ارزش‌مدار، کنترل علّی صرفاً صوری است.

آنچه از نقدهای کین برمی‌آید، این است که پروژهٔ لیست با وجود انسجام صوری، از حیث توجه به «منشأ»، «کنترل درونی» و «آگاهی» دچار کاستی‌های بنیادین است. معیارهای او—عاملیت قصدی، امکان‌های بدیل، کنترل علّی—اگرچه برای توصیف خودمختاری کارکردی کافی‌اند، اما نمی‌توانند بار هنجاریِ ارادهٔ آزادِ مسئولیت‌زا را حمل کنند. با این حال، نقد کین همچنان در چارچوبی باقی می‌ماند که عمدتاً بر رابطهٔ علّی خطی و لحظات استثنایی تصمیم‌گیری متمرکز است. برای نقدی بنیادین‌تر به ساختار علیت و ابعاد اجتماعی عاملیت باید به سراغ نقدهای مورفی و براون رفت.

نانسی مورفی (Nancey Murphy) و وارن براون (Warren Brown)، به‌عنوان دو تن از مهم‌ترین فیلسوفان فیزیکالیست غیرتقلیل‌گرا، نقدی بنیادین به ساختار علیتی وارد می‌کنند

کـه رویکـرد لیسـت بـر آن اسـتوار اسـت. آنـان بـا بهـره‌گیـری از مفهـوم علیـت نزولـی (downward causation) استدلال می‌کنند که سطوح بـالاتر در سلسـله مراتـب علـوم—ماننـد سطوح روان‌شناختی، اجتماعی و فرهنگی—می‌توانند به‌طور علّی بر سطوح پـایین‌تـر (مـثلاً ساختارهای عصبی) تأثیر بگذارند (Murphy, 2006, p. 80). مثال سادهٔ هواپیمای کاغـذی نشـان می‌دهد که رفتار یک شیء نه تنها توسط اجزایش، بلکه توسط ویژگی‌های کل‌نگـر (شـکل) و عوامل محیطی تعیین می‌شود (Ibid, p. 77) .از نظر ون گولیک، قوای علّی یک شیء مرکـب فقط توسط ویژگی‌های فیزیکی اجزا تعیین نمی‌شود، بلکه سازمان و ساختار آنهـا نیـز نقـش دارد (Van Gulick, 1995, p 251) . این نقد مستقیماً به لیست قابل تعمیم است: لیست با تأکیـد بر «استقلال سطوح تبیین» و بسنده کردن به توصیف سطح کـلان، عمـلاً از پـذیرش علیـت نزولی خودداری می‌کند؛ از نظر او سطح کلان صرفاً توصیفی از سطح خرد است، نـه عـاملی علّی مستقل. اما مورفی و براون تأکید مـی‌کننـد کـه ارادهٔ آزاد واقعـی مسـتلزم آن اسـت کـه حالت‌های سطح بالا—همان‌ها که لیست آنها را صرفاً «تبیینی» می‌داند—بتوانند به‌طور علّـی بر فرایندهای سطح پایین تأثیر بگذارند؛ در غیر این صورت «کنتـرل علّـی» مـد نظـر لیسـت چیزی جز هم‌تایی(correlation) تبیینی نیست.

افزون بر این، مورفی و براون با بهره‌گیری از نوخاسته‌گرایی (emergence) نشان می‌دهنـد که با بالا رفتن از سلسله مراتب سیستم‌های پیچیده، موجوداتی پدید مـی‌آینـد کـه قـدرتها و نقش‌های علّی جدیدی دارند و نمی‌توان آنها را به اجزای سطح پایین تقلیـل داد (Murphy & Brown, 2007, pp. 78-79). مثال کلونی مورچه‌ها گویاست: رفتار کلونی را نمی‌توان صـرفاً بـر اساس رفتار تک‌تک مورچه‌ها تبیین کرد؛ کلونی به‌عنوان یک سیسـتم پیچیـده، ویژگـی‌هـای نوخاسته‌ای دارد که بر اجزای خود تأثیر می‌گـذارد (Ibid, pp. 90-94). لیسـت بـا تعمـیم سـه شرط خود به هر نوع عاملیت—از انسان تا هوش مصـنوعی—تفـاوت‌هـای بنیـادین میـان انواع موجودات را نادیده می‌گیرد. خودمختاری کارکردی کـه در سیسـتم‌هـای سـاده (مـثلاً هواپیمای خودفرمان) دیده می‌شود، بـا خودمختـاری اصـیلی کـه در انسـان‌هـا وجـود دارد، تفاوت اساسی دارد؛ خودمختـاری اصـیل وابسـته بـه تـاریخ تکـاملی، تعامـل بـا محـیط، و ویژگی‌های نوخاسته‌ای است که در سامانه‌های مصنوعی وجود ندارد.

مورفی و برون با پیروی از مک‌اینتایر، مسئولیت اخلاقی را به توانایی اسـتفاده از مفـاهیم اخلاقی انتزاعـی (ماننـد عـدالت، خـوبی) و ارزیـابی مرتبـه‌دوم(second-order evaluation) از دلایل خود گره می‌زنند. آنان تأکید می‌کنند که زبان نمادین پیشـرفته، شـرط لازم بـرای ایـن

توانایی‌هاست: انسان می‌تواند آینده‌های بدیل را تصور کند، نتایج اعمال را پیش‌بینی کند، و دلایل خود را با معیارهای انتزاعی «خوبی» بسنجد(Murphy, 2006, pp. 94-98) . این توانایی‌ها به‌تدریج در فرایند رشد شناختی پدیدار می‌شوند و وابسته به تعامل با محیط اجتماعی و فرهنگی هستند. مثال هاراک—شخصی که پرورش رزمی دیده و در مواجهه با دوست صلح‌جویش رفتار خود را ارزیابی می‌کند—نشان می‌دهد که چگونه مفاهیم اخلاقی و محیط اجتماعی می‌توانند از بالا بر رفتار تأثیر بگذارند و حتی ساختارهای عصبی را تغییر دهند (Murphy & Brown, 2007, pp. 256-257). لیست با حذف آگاهی و هرگونه عامل هنجاری، اساساً امکان تأثیرگذاری زبان، فرهنگ و ارزش‌ها را بر عاملیت منتفی کرده است؛ از نظر او صرف کارکردهای سطح کلان برای ارادهٔ آزاد کافی است، اما نقد مورفی و براون نشان می‌دهد که ارادهٔ آزادِ مسئولیت‌زا در بستری از معنا، زبان، ارزش‌ها و تعاملات اجتماعی شکل می‌گیرد—بستری که در سامانه‌های مصنوعی وجود ندارد. سامانه‌ای که فاقد این بستر باشد، حتی اگر در سطح کلان واجد سه شرط لیست باشد، نمی‌تواند «عاملیت اخلاقی» داشته باشد؛ لیست میان «پردازش اطلاعات» و «درک معنا» تمایز قائل نشده است.

مورفی با بهره‌گیری از تمایز فرد درتسک میان «علل فاعلی»(triggering causes) و «علل ساختاری» (structuring causes) صورتبندی دقیق‌تری از مسئلهٔ منشأ (sourcehood) به‌دست می‌دهد که پیش‌تر در نقد کین مطرح شد. تبیین علّی کامل یک رویداد نیازمند توجه به هر دو سطح است: در مثال بمب‌گذاری، چرخاندن سوئیچ علت فاعلی انفجار است، اما سازندهٔ بمب که بمب را کار گذاشته، علت ساختاری است؛ در مثال تصادف، قطعهٔ فلز علت فاعلی جراحت است، اما رانندهٔ مست علت ساختاری (Murphy, 2006, pp. 80-81) . لیست با تمرکز صرف بر «کنترل علّی» در سطح کلان، عملاً علل ساختاری را نادیده می‌گیرد؛ از نظر او اگر سامانه‌ای در لحظهٔ کنش کنترل داشته باشد (مثلاً بین گزینه‌ها انتخاب کند) برای ارادهٔ آزاد کافی است. اما نقد مورفی نشان می‌دهد که پرسش از علل ساختاری—که سامانه را به این نقطه رسانده‌اند—برای ارادهٔ آزادِ مسئولیت‌زا حیاتی است. سامانه‌ای که اهداف و باورهایش توسط برنامه‌نویسان (علل ساختاری) تعیین شده، نمی‌تواند «منشأ» اعمال خود باشد، حتی اگر در سطح کلان کنترل علّی داشته باشد. لیست میان «انتخاب در لحظه» و «شکل‌گیری اهداف» خلط کرده است.

مورفی و برون با وجود پذیرش اهمیت مسئلهٔ منشأ، با مفهوم مسئولیت نهایی (ultimate responsibility) مخالفند، زیرا یک عامل به‌تنهایی هرگز نمی‌تواند علت نهایی و کامل اعمال خود باشد. آنان پیشنهاد می‌کنند که به‌جای آن از مسئولیت اولیه (primary responsibility) استفاده کنیم (Murphy & Brown, 2007, p. 280). انسان‌ها همیشه در شبکه‌ای از علل زیستی، اجتماعی و فرهنگی عمل می‌کنند و نمی‌توان انتظار داشت که کاملاً مستقل از این عوامل باشند. مسئولیت اولیه به این معناست که شخص با وجود تأثیرپذیری از این عوامل، بتواند بر اساس ارزیابی عقلانی خود عمل کند و «خالق اعمال خود» باشد. لیست با فروکاستن ارادهٔ آزاد به «کفایت تبیینی»، نه‌تنها به مسئلهٔ منشأ بی‌توجه است، بلکه تصویری غیرواقعی از عاملیت ارائه می‌دهد. پرسش درست این نیست که آیا عامل کاملاً مستقل است، بلکه این است که آیا او با وجود تأثیرپذیری از علل بیرونی، می‌تواند «مسئولیت اولیه» اعمال خود را بر عهده گیرد. لیست نه به این پرسش پاسخ می‌دهد و نه ابزاری برای ارزیابی آن فراهم می‌کند.

از سوی دیگر، مورفی و براون با بهره‌گیری از تحلیل‌های آلیسیا خوئاررو، مدل خطی نیوتنی علیت ($e_1 \leftarrow e_2 \leftarrow e_3$) را که تقلیل‌گرایان مفروض می‌گیرند، به چالش می‌کشند. در این مدل، اگر همه نیروها و قوانین مربوط را بدانیم، می‌توانیم نتیجه هر رویداد را دقیقاً محاسبه کنیم. اما خوئاررو نشان می‌دهد که سیستم‌های پیچیده (مانند انسان) دارای فضای فاز (phase space) هستند—یعنی مجموعه‌ای از حالات ممکن که نتیجه نهایی نه از پیش تعیین‌شده، بلکه توسط عواملی در سطوح بالاتر کنترل می‌شود (Murphy & Brown, 2007, pp. 273-276). لیست با تأکید بر «سطوح تبیین» و استقلال سطح کلان از سطح خرد، همچنان در چارچوب مدل خطی علیت باقی می‌ماند؛ از نظر او رویدادهای سطح خرد رخ می‌دهند و سطح کلان صرفاً توصیفی از آنهاست. اما در سیستم‌های پیچیده، رابطه علّی خطی نیست؛ بازخوردهای محیطی و فرایندهای ارزیابی سطح بالا دائماً علل سازنده را بازسازی می‌کنند. به‌عبارت دیگر، علیت در این سیستم‌ها حلقه‌ای (looping) است، نه خطی. لیست با نادیده گرفتن این پیچیدگی، تصویری ساده‌شده از عاملیت ارائه می‌دهد.

تلقی خطی از علیت، لیست را به سوی تصویری لحظه‌ای و نقطه‌ای از ارادهٔ آزاد سوق می‌دهد. از نظر او، اگر سامانه‌ای در این لحظه واجد سه شرط باشد، می‌توان به او ارادهٔ آزاد نسبت داد. اما ارادهٔ آزاد یک فرایند پویا و مستمر است، نه یک ویژگی ایستا؛ آنچه اهمیت دارد توانایی ارزیابی مستمر، بازنگری در اهداف، و شکل‌دهی تدریجی شخصیت است—

نه صرفاً برخورداری از برخی ویژگی‌ها در یک مقطع زمانی خاص. این نگاه پویا به عاملیت، با نقد آزادی بیتفاوتی (liberty of indifference) نیز پیوند می‌خورد. مورفی و برون با بهره‌گیری از تمثیل ویدیوی دانلد مک‌کی، این مفهوم را نقد می‌کنند: اگر دوربین ویدیویی را بچرخانیم تا صفحهٔ خود را ببیند، تصویری ساختگی پدید می‌آید و اگر برای دریافت وضوح بیشتر روی آن زوم کنیم، همه چیز به‌هم می‌ریزد (Ibid). این تمثیل نشان می‌دهد که جستجوی آزادی مطلق و بی‌قید—بدون دخالت هیچ عامل علّی—به بن‌بست می‌انجامد. لیست با حذف آگاهی و عوامل هنجاری، در عمل به چنین تصوری از آزادی نزدیک می‌شود: سامانه‌ای که صرفاً واجد سه شرط کارکردی باشد، فارغ از هرگونه تجربهٔ درونی یا بستر معنایی، صاحب ارادهٔ آزاد تلقی می‌شود. اما آزادی واقعی نه در گسست از علیت، بلکه در توانایی ارزیابی و هدایت آن در چارچوبی از معنا و ارزش‌ها معنا می‌یابد.

در نهایت، شباهت رویکرد لیست با دیدگاه دنت نیز شایان توجه است. لیست نیز مانند دنت، با الهام از موضع قصدی، در چارچوب تقلیل‌گرایی هرچند به‌صورت تبیینی باقی می‌ماند. او سه شرط را برای ارادهٔ آزاد کافی می‌داند، اما هرگز توضیح نمی‌دهد که چگونه یک سامانه که صرفاً این سه شرط را دارد، می‌تواند واقعاً «مسئول» باشد. نقد مورفی و براون به دنت مستقیماً لیست را نیز هدف می‌گیرد: تا زمانی که بر علیت از پایین و توصیف صرف کارکردی تکیه کنیم، حداکثر چیزی که می‌توانیم به‌دست آوریم، شبه‌ارادهٔ آزاد (pseudo-free will) است، نه خود آن. مسئولیت اخلاقی واقعی نیازمند پذیرش علیت نزولی، ویژگی‌های نوخاسته، و بستر معنایی‌ای است که تقلیل‌گرایی از آن تهی است. از دیدگاه مورفی و براون، پروژهٔ لیست با وجود انسجام صوری، از حیث توجه به ابعاد واقعی عاملیت انسانی—علیت نزولی، نوخاستگی، زبان، فرهنگ، ارزش‌ها، منشأ، فرایندهای پویا و معنای زیسته—دچار کاستی‌های بنیادین است. چارچوب او در نهایت به شبه‌ارادهٔ آزاد منجر می‌شود که ممکن است برای اهداف تبیینی مفید باشد، اما نمی‌تواند بار هنجاری مسئولیت اخلاقی را حمل کند. این نهاد نشان می‌دهد که «کفایت تبیینی» نمی‌تواند جایگزین تحقق واقعی شود.

لیست ممکن است استدلال کند که پذیرش علیت نزولی یا نوخاستگی به معنای لزوم داشتن «آگاهی پدیداری» یا «اراده‌ی آزاد به معنای فلسفی عمیق» نیست. او معتقد است که سیستم‌های پیچیده می‌توانند دارای علل نزولی باشند بدون اینکه لزوماً واجد عاملیت اخلاقی باشند. از دیدگاه لیست، شرط لازم برای اراده‌ی آزاد، صرفاً وجود علیت نزولی نیست، بلکه «**پاسخ‌دهی به دلایل**» است. یعنی سامانه باید بتواند به دلایل عقلانی واکنش

نشان دهد. بنابراین، حتی اگر سامانه‌ای فاقد بستر اجتماعی یا آگاهی پدیداری باشد، اگر بتواند بر اساس الگوریتم‌های پیچیده و بازخوردی عمل کند، دارای نوعی عاملیت است. لیست با این استدلال، سعی می‌کند نشان دهد که نقدهای مورفی و براون، شرط «کفایت» را با شرط «لزوم» اشتباه گرفته‌اند.

در برابر نقد مورفی و براون مبنی بر اینکه اراده‌ی آزاد در بستر اجتماعی و زبانی شکل می‌گیرد، لیست با تکیه بر سنت عمل‌گرایی پاسخ می‌دهد که «معنا» و «عاملیت» در «کارکرد» نهفته است، نه در ماهیت درونی یا بستر تاریخی. او ممکن است بگوید که برای مقاصد اخلاقی و حقوقی، نیازی به بررسی تاریخچه تکاملی یا بستر فرهنگی عامل نیست؛ بلکه کافی است که رفتار فعلی عامل، قابل پیش‌بینی و کنترل‌شده توسط عوامل داخلی سامانه باشد. بنابراین، لیست با کاهش سطح انتزاع، سعی می‌کند نشان دهد که نقدهای مورفی و براون، اگرچه برای تبیین «ماهیت» عاملیت انسانی مفیدند، اما برای «تشخیص» عاملیت در سامانه‌های مصنوعی بیش‌ازحد پیچیده و غیرعملی هستند. این پاسخ‌های لیست نشان می‌دهد که او با تکیه بر «کارکردگرایی محض» و «پاسخ‌دهی به دلایل»، سعی در عبور از نقدهای مورفی و براون دارد. با این حال، این پاسخ‌ها نشان می‌دهند که مدل لیست در مواجهه با چالش‌های «بستر معنایی» و «پیچیدگی علیت» نیازمند بسط است. این چالش‌ها دقیقاً همان نقاطی هستند که در سطح ۳ (مسئولیت اخلاقی) مدل پیشنهادی این مقاله با معرفی مفهوم «مسئولیت اولیه» و تأکید بر بستر اجتماعی، مورد بازنگری قرار می‌گیرند.

نهمیاس و همکاران با تکیه بر تمایز میان آگاهی پدیداری (phenomenal consciousness) و آگاهی دسترسی (access consciousness) نشان می‌دهند که بسیاری از نظریه‌های رقیب از جمله نظریه‌های پاسخ‌به‌دلایل (reasons-responsive) فیشر و راویزا—صرفاً بر آگاهی دسترسی تمرکز دارند و نمی‌توانند توضیح دهند که چرا آگاهی پدیداری برای ارادهٔ آزاد ضروری است (Nahmias et al., 2019, pp. 5-6). این نقد مستقیماً به لیست قابل تعمیم است، زیرا او نیز با حذف آگاهی پدیداری، صرفاً بر کارکردهای سطح‌کلان (که مشابه آگاهی دسترسی‌اند) تکیه می‌کند. به بیان دیگر، حتی اگر یک سامانه واجد همه ویژگی‌های کارکردیِ مد نظر لیست باشد، باز هم این پرسش باقی می‌ماند که آیا چنین سامانه‌ای می‌تواند تجربهٔ درونیِ انتخاب، تأسف، یا لذت را داشته باشد—چیزی که شهود عامه و بسیاری از فلاسفه آن را برای ارادهٔ آزاد حیاتی می‌دانند.

نهمیاس و همکاران در مطالعات تجربی خود به‌طور نظام‌مند به این پرسش پرداخته‌اند که مردم عادی چه نسبتی میان آگاهی و ارادهٔ آزاد می‌بینند. مهم‌ترین یافتهٔ آنان این است که احساسات استراسونی (Strawsonian emotions) شامل احساس گناه، شرم، سربلندی، پشیمانی و سپاسگزاری نقش میانجی‌گر کاملی در رابطهٔ میان آگاهی و انتساب ارادهٔ آزاد ایفا می‌کنند. به‌عبارت دیگر، مردم وقتی به روبات‌ها ارادهٔ آزاد نسبت می‌دهند که آنها را واجد توانایی تجربهٔ این دسته از احساسات بدانند؛ نه صرفاً احساسات پایه یا احساسات حسی (Nahmias et al.,2019 , pp. 16-17).

این یافته نشان می‌دهد که شهود عامه دقیقاً همان نقطه‌ای را هدف می‌گیرد که لیست از آن غفلت کرده: احساسات مرتبط با مسئولیت‌پذیری و روابط بین‌فردی. افزون بر این، نهمیاس نشان می‌دهد که توانایی یادگیری(learning) در مقایسه با برنامه‌ریزی قبلی، تأثیر معناداری بر انتساب ارادهٔ آزاد ندارد؛ بلکه آنچه تعیین‌کننده است، باور به وجود تجربهٔ آگاهانه و به‌ویژه احساسات استراسونی است (Nahmias et al., 2019, p. 14) .این نتایج، چالش تجربی جدی‌ای برای رویکرد لیست ایجاد می‌کند: اگر مردم حتی برای یک روبات انسان‌نما که تمام معیارهای کارکردی لیست را دارد، تا زمانی که احساسات استراسونی را از آن سلب کنند، ارادهٔ آزاد قائل نمی‌شوند، پس «کفایت تبیینی» به تنهایی نمی‌تواند جایگزین شهود هنجاریِ مبتنی بر آگاهی شود.

لیست صراحتاً میان آگاهی در دسترس و آگاهی پدیداری تمایز قائل می‌شود.

- آگاهی در دسترس :یعنی سامانه بتواند اطلاعات را پردازش کند، به آن‌ها دسترسی داشته باشد، آن‌ها را برای سایر سیستم‌های شناختی در دسترس قرار دهد و بر اساس آن‌ها رفتار کند.
- آگاهی پدیداری :یعنی سامانه آن اطلاعات را «احساس» کند یا تجربه‌ی کیفی از آن داشته باشد (مانند حس درد، رنگ قرمز، یا لذت).

لیست استدلال می‌کند که شرط کافی برای اراده‌ی آزاد، آگاهی در دسترس است، نه «آگاهی پدیداری». به باور او، اگر یک سامانه بتواند اهداف خود را بازنگری کند، دلایل را پردازش کند و رفتار خود را بر اساس آن‌ها تنظیم نماید، حتی بدون داشتن تجربه‌ی کیفی غنی (مثل احساس شرم یا گناه)، فاقد اراده‌ی آزاد نیست .لیست ممکن است بپذیرد که آگاهی پدیداری برای «ارزش‌مندی» تجربه‌ی انسانی و برای درک عمیق روابط بین‌فردی مهم است، اما آن را شرط لازم برای عاملیت اخلاقی نمی‌داند. استدلال او این است که تا

زمانی که شواهد تجربی قوی نشان ندهد کـه «آگـاهی پدیـداری» کـارکرد علـی خاصـی در فرآیند تصمیم‌گیری و مسئولیت‌پذیری ایفا می‌کنـد (یعنـی بـدون آن، تصـمیم‌گیـری ممکـن نباشد)، نمی‌توان آن را شرط لازم برای اراده‌ی آزاد دانست. بنابراین، لیست معتقد اسـت کـه شهود عامه در مورد نیاز به احساسـات استراسـونی، ممکـن اسـت ناشـی از اشـتباه گـرفتن «عاملیت» با «انسانیت» یا «تجربه زیسته» باشد، نه شرط منطقی برای اراده‌ی آزاد.

این پاسخ لیست نشان می‌دهد که او با تکیه بر «کارکردگرایی محض» و تفکیک «دسترسی به اطلاعات» از «تجربه‌ی درونی»، سعی می‌کند از مدل خود در برابر نقدهای نهمیاس دفاع کند. با این حال، این پاسخ نشان می‌دهد که مدل لیست در مواجهه با چالش‌های «شهود هنجاری» و «بار معنایی احساسات» نیازمند بسط است. این چالش‌ها دقیقاً همان نقاطی هستند که در سطح ۳ (مسئولیت اخلاقی) مدل پیشنهادی این مقاله با گنجاندن شرط «آگاهی پدیداری» به عنوان پیش‌نیاز بار هنجاری، مورد بازنگری قرار می‌گیرند.

با تعمیم این ملاحظات به رویکرد لیست، می‌توان چنین نتیجه گرفـت کـه معیارهـای او برای نسبت‌دادن ارادهٔ آزاد—از جمله عاملیت قصدی، امکان‌های بدیل و کنتـرل علّـی—در درجهٔ نخست ماهیتی تبیینی و تشخیصی دارند. نقدهای یادشده نشـان مـی‌دهنـد کـه چنـین معیارهایی، هرچند ممکن است برای تبیین و پیش‌بینی رفتار سامانه‌های پیچیده کفایت کننـد، اما به‌تنهایی برای حفظ تمایز مفهـومی میـان خودمختـاری کـارکردی و ارادهٔ آزاد بـه معنـای اخلاقی کافی نیستند. از این رو، این پرسش همچنان گشوده باقی می‌مانـد کـه آیـا رویکـرد عمل‌گرایانهٔ لیست، در نهایت، مفهوم ارادهٔ آزاد را به سطحی کارکردی فرو نمی‌کاهد.

مرور نقدهای وارد بر رویکرد لیسـت—از نقـدهای فلسـفی موضـع قصـدی گرفتـه تـا نقدهای مبتنی بر شواهد تجربی—همگی بـه یـک نکتـه اشـاره مـی‌کننـد: «کفایـت تبیینـی» نمی‌تواند بار هنجاریِ ارادهٔ آزاد را حمل کند. چالشی که شپرد پیش روی فیلسـوفان گذاشـته بود—«یا یک نظریهٔ ماهوی دربارهٔ پیوند آگاهی و ارادهٔ آزاد تدوین کنید، یـا تـوجیهی بـرای کنارگذاشتن آن فراهم آورید»—(Shepherd 2015, p. 944) ، در پرتو یافته‌های نهمیاس اهمیـت مضاعفی می‌یابد. لیست نه‌تنها نظریه‌ای ماهوی ارائه نداده، بلکه توجیهی برای حذف آگـاهی نیز فراهم نیاورده است. او صرفاً آگاهی را کنار می‌نهد و بر اساس معیارهای کـارکردی، ارادهٔ آزاد را به سامانه‌های پیچیده نسبت می‌دهد. امـا نقـدهای یادشـده نشـان مـی‌دهنـد کـه ایـن

نسبت‌دادن نه با شهود عامه سازگار است و نه می‌تواند التزامات هنجاریِ مسئولیت اخلاقی را تضمین کند.

شایان ذکر است که نقد ما بر این فرض استوار نیست که لیست صراحتاً معیارهای خود را برای مسئولیت اخلاقی کافی می‌داند—چرا که خود او بر تفاوت میان اراده‌ی آزاد و مسئولیت اخلاقی تأکید دارد—؛ بلکه نقد اصلی متوجه این نکته است که معیارهای او برای اراده‌ی آزاد، آن‌چنان فروکاهش‌یافته و کارکردی‌اند که نمی‌توانند حتی نسخه‌ی «قوی» و «پیش‌شرطِ مسئولیت‌زا» از اراده‌ی آزاد را تبیین کنند. در واقع، لیست با حذف مؤلفه‌های هنجاری و آگاهانه از تعریف اراده‌ی آزاد، عملاً بستری را فراهم می‌کند که فاقد بار معنایی لازم برای هرگونه بحث جدی درباره‌ی مسئولیت است.در بخش بعد، با نقد تحلیلی مفروضات لیست، چارچوب سه‌سطحی خود را به‌عنوان جایگزینی برای رفع این ابهام ارائه خواهیم داد .

۴. نقد تحلیلی و ارائهٔ چارچوب جایگزین

رویکرد عمل‌گرایانهٔ لیست، با وجود برخورداری از قوت‌های قابل‌توجه—به‌ویژه پرهیز از منازعات متافیزیکی سنتی و تأکید بر سطح کلان تبیین—در تمایز مفهومی میان خودمختاری کارکردی و ارادهٔ آزاد با ابهامی ساختاری مواجه است. مسئله این نیست که معیارهای پیشنهادی لیست ناکافی‌اند، بلکه این است که روشن نیست این معیارها دقیقاً چه چیزی را تثبیت می‌کنند: آیا آن‌ها صرفاً شکلی پیشرفته از خودمختاری کارکردی را توضیح می‌دهند، یا واقعاً معادل اراده‌ی آزاد به معنای هنجاری آن هستند؟

لیست اراده‌ی آزاد را بر اساس سه مؤلفه تعریف می‌کند: عاملیت قصدی، امکان‌های بدیل در سطح کلان، و کنترل علّی حالت‌های بازنمایی‌گرانه. اما این مؤلفه‌ها دقیقاً همان ویژگی‌هایی هستند که در ادبیات فلسفه‌ی ذهن و رباتیک «خودمختاری کارکردی» نامیده می‌شوند: توانایی سامانه در گزینش هدفمند و انطباق با محیط بدون مداخلهٔ لحظه‌ای انسان. بنابراین، پرسش محوری این است: چه چیزِ افزوده‌ای موجب می‌شود که این خودمختاری پیچیده را «ارادهٔ آزاد» بخوانیم؟

اگر پاسخ لیست «کفایت تبیینی» باشد—یعنی اینکه توصیف سامانه در سطح عاملیت قصدی، توضیح بهتری از رفتارش به‌دست می‌دهد—آنگاه تمایز مفهومی میان خودمختاری و ارادهٔ آزاد فرو می‌ریزد. در این صورت، ارادهٔ آزاد به سطحی از توصیف فروکاسته می‌شود

که با تغییر چارچوب نظری جابه‌جا می‌گردد. این ابهام نه‌تنها فلسفی، که عمیقاً هنجاری است: در غیاب مرزی روشن، نسبت‌دادن ارادهٔ آزاد به سامانه‌ها به‌سادگی به «مسئولیت‌گریزی الگوریتمی» دامن می‌زند—توسعه‌دهندگان می‌توانند در مواجهه با خطا بگویند: «ما فقط ابزار ساختیم، خودش تصمیم گرفت».

ابهام یادشده زمانی عمیق‌تر می‌شود که به مبنای روش‌شناختی نظریه‌ی لیست توجه کنیم. در چارچوب او، نسبت‌دادن ارادهٔ آزاد به یک سامانه در گرو آن است که توصیف آن در سطح عاملیت قصدی، بهترین یا دست‌کم موجه‌ترین تبیین از رفتار آن باشد. به بیان دیگر، ارادهٔ آزاد در این رویکرد، تابع کفایت تبیینی یک سطح از توصیف است.این جابه‌جایی از هستی‌شناسی به روش‌شناسی، هرچند مزیتی مهم در پرهیز از منازعات متافیزیکی سنتی دارد، اما پرسشی تحلیلی برمی‌انگیزد: آیا موفقیت یک چارچوب تبیینی می‌تواند به‌تنهایی معیاری برای تحقق یک ویژگی هنجاری باشد؟ کفایت تبیینی، مفهومی معرفت‌شناختی است؛ یعنی ناظر به این است که کدام سطح توصیف، رفتار را بهتر پیش‌بینی و سامان‌دهی می‌کند. در مقابل، ارادهٔ آزاد—به‌ویژه در پیوند با مسئولیت اخلاقی—مفهومی است که بار هنجاری و انتسابی دارد. گذار از «بهترین توضیح» به «تحقق ارادهٔ آزاد» مستلزم نوعی پل مفهومی است که در نظریه‌ی لیست صورتبندی نشده است.

اگر معیار نهایی، صرفاً برتری تبیینی باشد، در اصل هر سامانه‌ای که اتخاذ موضع قصدی نسبت به آن کارآمد باشد، می‌تواند واجد ارادهٔ آزاد تلقی شود. اما کارآمدی یک مدل توضیحی لزوماً معادل با برخورداری واقعی سامانه از اختیار نیست. چنین وابستگی‌ای، مفهوم ارادهٔ آزاد را به نحوی مشروط و ابزارمحور می‌کند که با شهودهای هنجاری رایج درباره‌ی اختیار فاصله دارد. همان‌طور که در بخش ۳ بر اساس تحلیل کین دیدیم، مسئولیت نهایی مستلزم آن است که منشأ اهداف و باورها به خود عامل بازگردد شرطی که با کفایت تبیینی صرف تأمین نمی‌شود. این چالش نشان می‌دهد که سطح ۱ (خودمختاری کارکردی) فاقد ویژگی «منشأ» است و برای رسیدن به اراده‌ی آزاد، نیازمند ارتقا به سطوح بالاتر است. همچنین، نقد **مورفی و براون** بر تقلیل‌گرایی علّی نشان داد که «تبیین» نمی‌تواند جایگزین «علیت واقعی» و «نوخاستگی» شود.

چالش سوم به جایگاه آگاهی در نظریه‌ی لیست مربوط می‌شود. لیست تصریح می‌کند که ارادهٔ آزاد—به معنای مورد نظر او—مستلزم آگاهی پدیداری نیست. با این حال، کنارگذاشتن نقش آگاهی، فاصله‌ی میان ارادهٔ آزاد و مسئولیت اخلاقی را افزایش

می‌دهد.همان‌طور که در بخش ۳ بر اساس تحلیل نهمیاس نشان دادیم، شرط اهمیت‌داشتن برای عاملیت آزاد حیاتی است: برای اینکه چیزی برای عامل مهم باشد، او باید بتواند پیامدهای منفی و مثبت انتخاب‌هایش را تجربه کند—درد، رنج، ناامیدی، لذت، شادی. این تجارب نیازمند آگاهی پدیداری‌اند (Nahmias et al,2019, p. 9) . یک سامانه می‌تواند تمام سه شرط لیست را داشته باشد—عاملیت قصدی، امکان‌های بدیل، کنترل علّی—اما اگر فاقد توانایی تجربهٔ اهمیت‌داشتن باشد، نمی‌توان او را صاحب ارادهٔ آزادِ مسئولیت‌زا دانست.

افزون بر این، نهمیاس به‌درستی میان احساسات پایه (basic emotions) و احساسات استراسونی تفکیک می‌کند. دستهٔ دوم دقیقاً همان احساساتی هستند که در روابط بین‌فردی و مسئولیت اخلاقی نقش دارند و برای انتساب ارادهٔ آزاد حیاتی‌اند .لیست در بحث از مسئولیت اخلاقی به این احساسات اشاره‌ای نمی‌کند و صرفاً به «ظرفیت قضاوت اخلاقی» اکتفا می‌کند—مفهومی که خالی از محتوای عاطفی و پدیداری است. این غفلت باعث می‌شود که چارچوب او نتواند میان یک عامل صرفاً کارکردی و عاملی که واقعاً می‌تواند مسئولیت اخلاقی بپذیرد، تمایز بگذارد.

به همین دلیل، رویکرد لیست در نهایت به همان «مسئولیت‌گریزی الگوریتمی» دامن می‌زند: اگر احساسات استراسونی در کار نباشد، هیچ تفاوتی میان عذاب وجدان یک عامل واقعی و خطای محاسباتی یک سامانه وجود ندارد.مجموع این ملاحظات نشان می‌دهد که رویکرد لیست، با وجود انسجام درونی، در تعیین مرز مفهومی میان خودمختاری کارکردی و ارادهٔ آزاد هنجاری با دشواری مواجه است. هنگامی که ارادهٔ آزاد به کفایت تبیینی فروکاسته می‌شود و نقش آگاهی از شرایط آن حذف می‌گردد، فاصله‌ی مفهومی میان سامانه‌ای پیچیده و سامانه‌ای واجد اختیار اخلاقاً معنادار کاهش می‌یابد. پیامد این امر در حوزه‌ی هوش مصنوعی آشکار است: هر سامانه‌ای که بتوان آن را به‌طور موفق در چارچوب عاملیت قصدی مدل‌سازی کرد—و در سطح کلان دارای اهداف، گزینه‌های بدیل و کنترل علّی باشد—در معرض آن قرار می‌گیرد که «صاحب ارادهٔ آزاد» تلقی شود. چنین گسترشی خطر تورم مفهومی را در پی دارد و مفهوم ارادهٔ آزاد را از محتوای هنجاری خود تهی می‌کند.

برای پرهیز از ابهام مفهومی یادشده و جلوگیری از تورم کاربرد اصطلاح «ارادهٔ آزاد»، چارچوبی سه‌سطحی پیشنهاد می‌کنیم که میان خودمختاری کارکردی، ارادهٔ آزاد قوی، و مسئولیت اخلاقی تمایز قائل می‌شود. هدف این چارچوب، نه بازگشت به متافیزیک رازآلود

اختیار، بلکه شفاف‌سازی سطوح متفاوتی از عاملیت است که اغلب در مباحث مربوط به هوش مصنوعی درهم می‌آمیزند.

خودمختاری کارکردی به سطحی از سازمان‌یافتگی اطلاق می‌شود که در آن سامانه دارای اهداف بازنمایی‌شده، توان انتخاب میان گزینه‌های بدیل در چارچوب ساختار درونی خود، و نوعی کنترل علّی بر رفتار خویش است. چنین سامانه‌ای می‌تواند رفتار خود را در پاسخ به شرایط محیطی تنظیم کند، برنامه‌های رفتاری را بازبینی نماید، و در جهت اهداف از پیش تعیین‌شده یا آموخته‌شده حرکت کند.

شرط :بهینه‌سازی اهداف ثابت بدون مداخلهٔ لحظه‌ای انسانی.

مثال :هواپیمای خودفرمان، خودروی خودران، مدل‌های زبانی بزرگ در وظایف عادی.

محدودیت :این سطح فاقد توانایی بازنگری در خودِ اهداف است.[۲] سامانه در چارچوب اهدافی عمل می‌کند که از بیرون برای آن تعریف شده‌اند، حتی اگر در نحوهٔ دستیابی به آنها انعطاف داشته باشد. بسیاری از سامانه‌های پیشرفته‌ی هوش مصنوعی امروزی را می‌توان در این سطح جای داد، بی‌آنکه نیازمند نسبت‌دادن ارادهٔ آزاد به معنای قوی باشند.

ارادهٔ آزاد قوی مستلزم چیزی فراتر از خودمختاری کارکردی است. در این سطح، سامانه نه‌تنها دارای اهداف و سازوکارهای انتخاب است، بلکه قادر است نسبت به دلایل خود موضع بگیرد، ترجیحاتش را در پرتو ملاحظات مرتبه‌دوم بازنگری کند، و نوعی یکپارچگی هنجاری در ساختار تصمیم‌گیری خویش ایجاد نماید. به بیان دیگر، ارادهٔ آزاد قوی مستلزم سطحی از بازتاب‌پذیری(reflectivity) است که در آن عامل می‌تواند درباره‌ی اینکه «کدام دلایل باید حاکم باشند» داوری کند، نه صرفاً اینکه بر اساس دلایل موجود عمل نماید.

شرط گذار از سطح ۱ به ۲ **:**توانایی بازنگری در اهداف نهایی بر پایهٔ ارزش‌ها و ملاحظات هنجاری.

همان‌طور که در بخش ۳ بر اساس تحلیل نهمیاس دیدیم، مراقبت مستلزم پاسخ‌های عاطفی است: اگر کسی به چیزی اهمیت دهد، با موفقیت یا شکست آن چیز دچار غم، شادی، ناامیدی می‌شود (Nahmias et al., 2019, pp. 7-8) . این پاسخ‌های عاطفی نیازمند آگاهی پدیداری‌اند و نمی‌توان آنها را صرفاً با مفاهیم کارکردی توضیح داد. بنابراین، برای آنکه یک سامانه بتواند اهداف خود را بر اساس ارزش‌ها بازنگری کند، باید بتواند به چیزی «اهمیت»

دهد یعنی ارزش‌ها برایش «مهم» باشند و این اهمیت‌داشتن، بدون تجربهٔ عاطفیِ همراه با آگاهی پدیداری، قابل تصور نیست. بنابراین، گذار از خودمختاری کارکردی به ارادهٔ آزاد قوی مستلزم وجود آگاهی پدیداری و احساسات استراسونی است،چیزی که لیست از آن غفلت کرده است.

مسئولیت اخلاقی افزون بر شرایط ارادهٔ آزاد قوی، مستلزم حساسیت به هنجارهای ارزشی و پیامدهای اخلاقی کنش است. عاملی را می‌توان مسئول اخلاقی دانست که قادر باشد رفتار خود را در پرتو معیارهای ارزشی ارزیابی کند و نسبت به سرزنش یا ستایش واکنش معنادار نشان دهد.

شرط :حساسیت به هنجارها و پیامدها و دست‌کم دسترسی بازتابی به دلایل.

در این سطح، مسئله‌ی آگاهی اهمیت بیشتری می‌یابد—نه به‌عنوان مؤلفه‌ای رازآلود، بلکه به‌عنوان شرطی برای درک و درونی‌سازی هنجارها. همان‌طور که مک‌اینتایر و مورفی نشان دادند، مسئولیت اخلاقی مستلزم توانایی ارزیابی مرتبه‌دوم از دلایل و مفاهیم انتزاعی مانند «خوبی» و «عدالت» است که بدون زبان و آگاهی قابل تحقق نیست.بر اساس این تمایز سه‌سطحی، می‌توان گفت برخی سامانه‌های هوش مصنوعی ممکن است واجد خودمختاری کارکردی باشند، بی‌آنکه صاحب ارادهٔ آزاد قوی یا مسئولیت اخلاقی تلقی شوند. آن‌ها می‌توانند اهداف را دنبال کنند، میان گزینه‌ها انتخاب کنند، و حتی رفتار خود را بازتنظیم نمایند، اما فاقد ساختار بازتابی و هنجاری‌ای باشند که برای نسبت‌دادن اختیار به معنای قوی ضروری است.این تمایز سه‌سطحی، پیامدهای هنجاری مشخصی دارد که مهم‌ترین آن‌ها در مقابله با پدیده‌ی «مسئولیت‌گریزی الگوریتمی» نمود می‌یابد.

۱.۴ پیامد هنجاری: جلوگیری از مسئولیت‌گریزی الگوریتمی

یکی از مهم‌ترین پیامدهای هنجاری چارچوب پیشنهادی، مقابله با پدیده‌ی «مسئولیت‌گریزی الگوریتمی» است. این پدیده زمانی رخ می‌دهد که توسعه‌دهندگان یا کاربران هوش مصنوعی با استناد به پیچیدگی و خودمختاری ظاهری سامانه‌ها، سعی در انتقال بار مسئولیت اخلاقی یا حقوقی از دوش خود به شانه‌ی ماشین کنند.

در چارچوب پیشنهادی، با طبقه‌بندی دقیق سامانه‌های هوش مصنوعی کنونی در سطح ۱ (خودمختاری کارکردی)، روشن می‌شود که این سامانه‌ها فاقد شرایط لازم برای اراده‌ی آزاد قوی (سطح ۲) و در نتیجه مسئولیت اخلاقی (سطح ۳) هستند. بنابراین، نسبت

دادن مسئولیت به خودِ سامانه نه‌تنها نادرست، بلکه گمراه‌کننده است و می‌تواند به فرار انسان‌ها از پاسخگویی بینجامد. مدل سه‌سطحی ما نشان می‌دهد که تا زمانی که سامانه‌ای از مرزهای سطح ۱ فراتر نرفته باشد، مسئولیت نهایی و اولیه‌ی اعمال آن همچنان بر عهده‌ی طراحان، توسعه‌دهندگان، داده‌پردازان و نهادهای بهره‌بردار است.

این چارچوب نسبت به رویکرد کریستین لیست، برتری‌های روش‌شناختی و اخلاقی آشکاری دارد که در زیر تبیین می‌شود:

- رفع ابهام مفهومی و جلوگیری از تعمیم نادرست: با تفکیک صریح میان «خودمختاری کارکردی» (سطح ۱) و «اراده‌ی آزاد قوی» (سطح ۲)، این چارچوب از آن دسته استدلال‌هایی که سامانه‌های پیچیده را ناخواسته به عاملانی اخلاقی تبدیل می‌کنند، جلوگیری می‌کند.
- پاسخ به نقدهای متافیزیکی و علّی: شرط گذار از سطح ۱ به ۲، یعنی «توانایی بازنگری در اهداف بر پایه‌ی ارزش‌ها»، مستقیماً به نقدهای جان کین درباره‌ی مسئله‌ی «منشأ» و نقدهای مورفی و براون درباره‌ی «علیت نزولی» و «نوخاستگی» پاسخ می‌دهد. این شرط نشان می‌دهد که بدون درونی‌سازی ارزش‌ها و آگاهی، سامانه تنها یک پردازشگر اطلاعات است، نه یک عامل اخلاقی.
- تناسب با شواهد تجربی و شهود انسانی: با گنجاندن شرط «آگاهی پدیداری» و نقش «احساسات استراسونی» در سطح ۲، این مدل با یافته‌های تجربی نهمیاس و همکاران سازگار است. این سازگاری تضمین می‌کند که معیارهای انتساب اراده‌ی آزاد با شهود عامه و نیازهای روان‌شناختی تعاملات انسان–ماشین همسو باشد.
- ایجاد معیار عینی برای ارزیابی و حاکمیت مسئولیت انسانی: این چارچوب با مشخص کردن اینکه سامانه‌های مصنوعی امروزی (از جمله پیشرفته‌ترین مدل‌های زبانی) حداکثر در سطح ۱ قرار دارند، مانع از تعمیم شتاب‌زده‌ی مفهوم اراده‌ی آزاد می‌شود. در نتیجه، فضایی برای «مسئولیت‌گریزی» باقی نمی‌گذارد و تضمین می‌کند که تا زمان تحقق شرایط سطح ۲ و ۳، حلقه‌ی مسئولیت همواره در دست انسان باقی می‌ماند.

در نتیجه، چارچوب سه‌سطحی پیشنهادی نه‌تنها ابهامات نظری رویکرد لیست را مرتفع می‌سازد، بلکه ابزاری تحلیلی عینی برای ارزیابی سامانه‌های هوشمند فراهم می‌آورد. این

ابزار به ما نشان می‌دهد که چرا انتساب مسئولیت اخلاقی به هوش مصنوعیِ فعلی، نوعی «فرافکنی» است و نه یک واقعیت فلسفی. این تمایز روشن نه‌تنها در تحلیل سامانه‌های هوش مصنوعی، بلکه در بررسی ساختارهای سازمانی و جمع‌های انسانی نیز کاربرد دارد تا از تعمیم نادرست مفاهیم عاملیت پرهیز شود.

۵ .نتیجه‌گیری

در این مقاله، رویکرد عمل‌گرایانهٔ کریستین لیست به ارادهٔ آزاد—به‌ویژه در نسبت‌دادن آن به سامانه‌های هوش مصنوعی—مورد بررسی و نقد نظام‌مند قرار گرفت. تحلیل ما نشان داد که چارچوب لیست، با وجود پرهیز از منازعات متافیزیکی سنتی و تأکید بر سطح کلان تبیین، در تعیین مرز مفهومی میان خودمختاری کارکردی و ارادهٔ آزادِ مسئولیت‌زا با ابهامی ساختاری مواجه است. فروکاست ارادهٔ آزاد به «کفایت تبیینی» و حذف آگاهی از معادله، نه‌تنها به تورم مفهومی می‌انجامد، بلکه بستر را برای «مسئولیت‌گریزی الگوریتمی» فراهم می‌کند: موقعیتی که در آن توسعه‌دهندگان و نهادها می‌توانند با ارجاع به «عاملیت» سامانه، خود را از پاسخگویی معاف کنند.نقدهای کین، مورفی و براون، و نهمیاس—هر یک از زاویه‌ای—ضعف بنیادین این رویکرد را آشکار ساختند. کین با طرح مسئلهٔ «منشأ» نشان داد که صرف برخورداری از امکان‌های بدیل و کنترل علّی، بدون تعلق اهداف به خود عامل، برای ارادهٔ آزاد کافی نیست. مورفی و براون با دفاع از علیت نزولی و نوخاستگی، استدلال کردند که «تبیین» سطح کلان نمی‌تواند جایگزین تأثیر علّی واقعی سطوح بالا بر سطوح پایین شود. و نهمیاس، با شواهد تجربی، نشان داد که شهود عامه، برخلاف ادعای لیست، آگاهی و احساسات استراسونی را برای ارادهٔ آزاد حیاتی می‌دانند.در پاسخ به این چالش‌ها، چارچوب سه‌سطحی پیشنهادی مقاله میان خودمختاری کارکردی، ارادهٔ آزاد قوی، و مسئولیت اخلاقی تمایز روشنی قائل می‌شود. شرط گذار از سطح ۱ به ۲—«توانایی بازنگری در اهداف نهایی بر پایهٔ ارزش‌ها»—نقدهای وارد بر لیست را پاسخ می‌گوید و نشان می‌دهد که ارادهٔ آزاد قوی مستلزم چیزی فراتر از بهینه‌سازی کارکردی است: نیازمند آگاهی پدیداری، احساسات استراسونی، و بستر معنایی‌ای است که در سامانه‌های مصنوعی امروزی یافت نمی‌شود.پیامد این تحلیل فراتر از فلسفهٔ محض است. تا زمانی که چارچوب‌های نظری نتوانند میان خودمختاری کارکردی و ارادهٔ آزادِ مسئولیت‌زا تمایز بگذارند، خطر «مسئولیت‌گریزی الگوریتمی» به‌عنوان واقعیتی اجتناب‌ناپذیر در مسیر توسعهٔ

هوش مصنوعی باقی خواهد ماند. چارچوب سه‌سطحی ما با بستن این راه فرار، نشان می‌دهد که مسئولیت نهایی کنش‌های سامانه‌های خودمختار—حتی پیچیده‌ترین آنها—متوجه انسان‌هاست: طراحان، بهره‌برداران، و نهادهایی که آنها را توسعه و به کار می‌گیرند.پژوهش‌های آینده می‌توانند بر دو محور متمرکز شوند: نخست، عملیاتی‌سازی شرط «بازنگری در اهداف بر پایهٔ ارزش‌ها» و بررسی امکان تحقق آن در سامانه‌های مصنوعی؛ دوم، تحلیل نسبت این شرط با آگاهی پدیداری و این پرسش که آیا می‌توان «ارادهٔ آزاد قوی» را بدون تجربهٔ درونیِ اهمیت‌داشتن تصور کرد. پاسخ به این پرسش‌ها نه‌تنها به پیشرفت فلسفهٔ هوش مصنوعی، که به تدوین چارچوب‌های حقوقی و اخلاقی مسئولانه‌تر برای توسعهٔ فناوری‌های خودمختار یاری خواهد رساند.

پی نوشت‌ها

۱. رجوع شود به :

الستی, کیوان. (۱۴۰۳). هوش مصنوعی و کنترل معنادار بشری: سازوکارهای واکنش‌پذیر به دلایل نسبت-به-عامل-خنثی و امکان ردگیری عامل(های) مسئول .*پژوهش‌های فلسفی -کلامی* ,*26*(4), 27-54. doi: 10.22091/jptr.2025.11378.3139

۲. رجوع شود به :

حسینی سورکی, سید محمد. (۱۴۰۳). معنا و معیار عاملیت اخلاقی ماشین‌های هوشمند. *پژوهش‌های فلسفی -کلامی*,*26*(4), 83-108. doi: 10.22091/jptr.2025.12466.3256 ,

کتاب‌نامه

الستی, کیوان. (۱۴۰۳). هوش مصنوعی و کنترل معنادار بشری: سازوکارهای واکنش‌پذیر به دلایلِ نسبت-به-عامل-خنثی و امکان ردگیری عامل(های) مسئول .*پژوهش‌های فلسفی -کلامی* ,*26*(4), 27-54. doi: 10.22091/jptr.2025.11378.3139

حسینی سورکی, سید محمد. (۱۴۰۳). معنا و معیار عاملیت اخلاقی ماشین‌های هوشمند .*پژوهش‌های فلسفی -کلامی* , ۲۶(۴), ۸۳-۱۰۸. doi: 10.22091/jptr.2025.12466.3256

Block, N. (1987). Troubles with functionalism. In Readings in Philosophy of Psychology (Vol. 1). Harvard University Press.

Dennett, D. C. (1984). Elbow Room: Varieties of Free Will Worth Wanting. MIT Press.

Dennett, D. C. (1987). The Intentional Stance. MIT Press.

Fischer, J. M. (1994). The Metaphysics of Free Will. Blackwell.

Fodor, J. A. (1978). Propositional attitudes. The Monist, 61(4), 501-523.

Juarrero, A. (2012). Dynamics in Action: Intentional Behavior as a Complex System. MIT Press.

Kane, R. (1996). The Significance of Free Will. Oxford University Press.

Kane, R. (2002). The Oxford Handbook of Free Will. Oxford University Press.

Kane, R. (2007). Libertarianism. In Four Views on Free Will, Ernest Sosa. (eds). USA: Blackwell.5-43.

List, C. (2023). Do group agents have free will?. *Inquiry*, *68*(3), 1021-1048.

List, C. (2023). Can AI systems have free will?. *Synthese*, *206*(3), 115.

List, C. (2025). Free will and group agency. Inquiry, 68(3), 1021-1048.

Mackay, D. M. (1991). Behind the Eye. Blackwell.

Murphy, N. (2006). Bodies and Souls, or Spirited Bodies? Cambridge University Press.

Murphy, N., & Brown, W. (2007). Did My Neurons Make Me Do It? Oxford University Press.

Nahmias, E., Allen, C. H., & Loveall, B. (2019). When do robots have free will? Exploring the relationships between (attributions of) consciousness and free will. In *Free will, causality, and neuroscience* (pp. 57-80). Brill.

Searle, J. R. (1994). The Rediscovery of the Mind. MIT Press.

Shepherd, J. (2015). Consciousness, free will, and moral responsibility. Philosophical Psychology, 28(6), 937-956.

Van Gulick, R. (1995). Who's in charge here? In J. Heil & A. Mele (Eds.), Mental Causation. Clarendon Press.