Shiraz University
RICeST
ISC
ISSN: 2008-7926

Journal of
Legal Studies
Scientific

Vol. 17, Issue 4, Winter 2026

JLS

Journal of Legal Studies
Journal Homepage: https://jls.shirazu.ac.ir/
doi: https://10.22099/JLS.2026.55440.5456

Research Article

Autonomous Weapons Based on Artificial Intelligence and Human Control: Challenges and Solutions in International Humanitarian Law

Masoumeh Zamanian[1*], Zahra Vatani[2]

1. Master's degree in International Law, Payam Noor Law School, Tehran, Iran;Graduated from the Department of Islamic Jurisprudence and Law, Faculty of Theology, Al-Zahra University, Tehran, Iran
2. Assistant Professor, Department of Jurisprudence and Law, Imam Khomeini and Islamic Revolution Research Institute, Tehran, Iran

Article history: Received: 06-12-2025 Accepted: 10-02-2026

Abstract

Introduction

Following the emergence of artificial intelligence on the global technological stage, its application in military weaponry—which, unlike traditional programming models, can learn how to perform a task through machine learning (i.e., training and post-use feedback)—poses a fundamental challenge to the principle of human control within International Humanitarian Law (IHL). IHL seeks to regulate the conduct of parties to armed conflict and mitigate unnecessary suffering (while acknowledging the military necessity of weakening the adversary). Since the inherent capacity for continuous "learning" from the environment renders the behavior of a machine learning system and its processing of new inputs unpredictable and inscrutable, the very possibility of meaningful human control over such systems, characterized by unpredictable behavior and opaque decision-making processes, is called

Please cite this article as:
Zamanian, M., Vatani, Z (2025). Autonomous Weapons Based on Artificial Intelligence and Human Control: Challenges and Solutions in International Humanitarian Law. *Journal of Legal Studies*, *17*(4), 313-368. doi: https://10.22099/JLS.2026.55440.5456

* Corresponding author:
E-mail address: m.zamanian@student.alzahra.ac.ir

into serious doubt. An even more profound question arises: are machine learning systems inherently dependent on human control to ensure compliance with IHL in the first place?

Methods

This article first provides a technological overview of autonomous learning systems, explicating their capabilities for continuous learning, analysis, and adaptation based on experience, alongside their opaque decision-making processes and unpredictable outcomes. This analysis establishes how these characteristics complicate the exercise of meaningful human control. Subsequently, by examining a system's capacity to rapidly process vast datasets and employ them in making strategic and precise decisions, the article demonstrates the significant potential of such systems to enhance overall compliance with IHL.

Results and Discussion

Machine learning is now deployed across a broad spectrum of military activities, primarily by reducing the rate of erroneous actions and reactions in complex environments. Relying on deep learning and reinforcement learning, these systems have transcended human limitations in solving complex battlefield problems. Unencumbered by human cognitive constraints and biases, they prove far more efficient at generating innovative solutions. However, this very capability renders machine behavior, both pre- and post-event, unpredictable and incomprehensible. This opacity may impede adherence to IHL's precautionary requirements, including the duty of constant care. Regarding the role of human control in fulfilling this duty, three principal theories are debated: "meaningful human control," "the necessity of human intervention prior to attack decisions," and "the possibility of eliminating prior human intervention."

Under the third theory, given that the decision-making of learning systems in sensitive situations is founded upon big data and processes opaque to human understanding, direct human supervision or intervention is rendered practically meaningless. Crucially, however, even absent human comprehensibility and direct control, such systems might be capable of making targeting decisions that are *more* compliant with IHL than those made by human-controlled systems. This potential superiority may stem from their ability to process and analyze immensely larger volumes of data with precision and speed exceeding human faculties. Consequently, their use should not be prohibited *solely* on the grounds of precluding meaningful human supervision, unless conclusive evidence is presented establishing the absolute necessity of such oversight.

Conclusion

The deployment of machine learning-based autonomous systems in combat reveals the imperative to redefine "meaningful human control." The findings indicate that under certain conditions, these systems can operate in accordance with IHL principles without direct human supervision, while in others, human control remains essential to guarantee compliance. The conventional wisdom that "more human control is invariably better" requires reassessment. The focus must shift toward ensuring "acceptable predictability" in system performance, such that it aligns with IHL standards and remains evaluable. This framework clarifies the pathways for system design, testing, and certification, applying human oversight at the most consequential stages.

Ultimately, through responsible design and deployment, learning systems can harness technological capabilities to reduce civilian casualties while upholding core IHL principles, such as military necessity and the mitigation of unnecessary harm. The proposed approach establishes an equilibrium between technological capability and human responsibility, enabling the optimal utilization of novel technologies within the bounds of International Humanitarian Law.

Keywords: Artificial Intelligence, Autonomous Weapons, International Humanitarian Law, Human Control, Machine Learning.

.

مجله مطالعات حقوقی
علمی
شاپا: ۷۹۲۶-۲۰۰۸

دوره هفدهم، شماره چهارم، زمستان 1404

مجله مطالعات حقوقی
Journal Homepage: https://jls.shirazu.ac.ir/
doi: https://10.22099/JLS.2026.55440.5456

JLS

مقاله پژوهشی

نقش سلاح‌های خودکار مبتنی بر هوش مصنوعی و کنترل انسانی در حقوق بین‌الملل بشردوستانه: چالش یا تعامل؟

معصومه زمانیان[1*]، زهرا وطنی[2]

1. دانش‌آموخته کارشناسی ارشد حقوق بین الملل، دانشکده حقوق پیام نور تهران، ایران؛ دانش آموخته کارشناسی ارشد، گروه فقه و حقوق اسلامی، دانشکده الهیات الزهراء، تهران، ایران
2. استادیار گروه فقه و حقوق پژوهشکده امام خمینی و انقلاب اسلامی، تهران، ایران

اطلاعات مقاله: تاریخ دریافت: 1404/09/15 تاریخ پذیرش: 1404/11/21

چکیده

مقدمه: پس از ظهور هوش مصنوعی در عرصهٔ فناوری جهانی، به‌کارگیری آن در تسلیحات نظامی که به جای تکیه بر مدل برنامه‌نویسی گذشته، می‌تواند از طریق یادگیری ماشینی یعنی آموزش و بازخورد پس از استفاده کاربر، یاد بگیرد چگونه وظیفه‌ای را به انجام رساند، علی‌رغم مزایای متعدد، کنترل انسانی را به‌عنوان اصل بنیادی حقوق بین‌الملل بشردوستانه که در پی تنظیم رفتار طرف‌های درگیر در مخاصمات مسلحانه و کاستن از رنج‌های غیرضروری (در عین توجه به ضرورت تضعیف دشمن) است به چالش می‌کشد. از آنجا که وجود قابلیت «یادگیری» دائمی از محیط در سامانه یادگیرنده ماشینی، رفتار آن و چگونگی پردازش ورودی‌های جدید را غیرقابل پیش‌بینی و فهم می‌کند، امکان کنترل‌گری انسانی برای چنین سامانه‌هایی با رفتار غیرقابل پیش‌بینی و فرآیندهای تصمیم‌گیری غیرشفاف محل پرسش جدی است و از آن بالاتر این سؤال مطرح می‌شود که آیا اساساً سیستم‌های یادگیری ماشین برای رعایت IHL نیازمند کنترل انسانی هستند؟

استناد به این مقاله:

زمانیان، معصومه و وطنی، زهرا (1404). نقش سلاح‌های خودکار مبتنی بر هوش مصنوعی و کنترل انسانی در حقوق بین‌الملل بشردوستانه: چالش یا تعامل؟. *مجله مطالعات حقوقی*. شماره 17. (4).

.313-368

* نویسنده مسئول: E-mail address: m.zamanian@student.alzahra.ac.ir

روش‌ها: این مقاله ابتدا در ضمن مروری تکنولوژیک بر سامانه‌های یادگیرنده خودکار، به توضیح توانایی این سامانه‌ها در یادگیری و تحلیل و تطبیق مداوم تجربیات و فرآیندهای تصمیم‌گیری غیرشفاف و نتایج غیرقابل پیش‌بینی و در نتیجه دشوار شدن کنترل معنادار انسانی بر آن‌ها می‌پردازد و سپس با تحلیل قدرت سامانه در پردازش سریع حجم عظیم داده‌ها و به‌کارگیری آن‌ها در اتخاذ تصمیمات استراتژیک و دقیق، پتانسیل بالا برای افزایش رعایت IHL توسط این سامانه‌ها را اثبات خواهد کرد.

یافته‌ها: امروزه یادگیری ماشینی با کاهش نرخ کنش و واکنش‌های اشتباه در محیط‌های پیچیده، در طیف گسترده‌ای از فعالیت‌های نظامی به‌کار گرفته می‌شود. این سامانه‌ها با ابتنا بر دو نوع یادگیری عمیق و تقویتی از مرز توانایی‌های انسانی در حل مسائل پیچیده جنگی عبور کرده و به دلیل عدم محدودیت به ظرفیت‌ها و پیش‌فرض‌های انسانی، در ایجاد راه‌حل‌های نوآورانه بسیار کارآمدتر هستند که البته این ویژگی رفتار ماشین‌ها را در مراحل قبل و بعد از وقوع غیرقابل پیش‌بینی و غیرقابل فهم می‌سازد و ممکن است رعایت الزامات احتیاطی حقوق بین‌الملل بشردوستانه ازجمله تکلیف مراقبت دائمی را با دشواری مواجه سازد که در مورد نقش کنترل انسانی در ایفای این تکلیف، سه نظریه «کنترل انسانی معنادار»، «ضرورت دخالت انسانی پیش از تصمیم‌گیری در حملات» و «امکان حذف دخالت انسانی پیشینی» مطرح است. در نظریه سوم با توجه به ابتنای تصمیم‌گیری سامانه‌های یادگیرنده در شرایط حساس بر داده‌های کلان و فرآیندهای غیرقابل فهم انسانی باید گفت نظارت یا دخالت انسان عملاً بی‌معنی است و البته حتی با وجود غیرقابل‌فهم بودن و عدم کنترل انسانی، ممکن است این سامانه‌ها قادر باشند تصمیمات هدفگیری منطبق بر IHL را بهتر از سامانه‌های تحت کنترل انسانی اتخاذ کنند. این توانایی ممکن است ناشی از قابلیت پردازش و تحلیل حجم بسیار بیشتری از داده‌ها با دقت و سرعتی فراتر از انسان باشد. بنابراین، نباید صرفاً به دلیل عدم امکان نظارت معنادار انسانی، استفاده از آن‌ها ممنوع شود، مگر اینکه دلیلی قطعی بر ضرورت نظارت انسانی اقامه شود.

نتیجه‌گیری: پیاده‌سازی سامانه‌های خودکار مبتنی بر یادگیری ماشینی در میدان‌های نبرد، ضرورت بازتعریف مفهوم «کنترل انسانی معنادار» را آشکار می‌سازد. یافته‌ها نشان می‌دهد که این سامانه‌ها در برخی شرایط می‌توانند بدون نظارت مستقیم انسانی با اصول حقوق

بین‌الملل بشردوستانه (IHL) منطبق عمل کنند و در عین حال، در شرایط خاص، اعمال کنترل انسانی برای تضمین رعایت IHL ضروری است. رویکرد متداول «هرچه کنترل انسانی بیشتر، بهتر» نیازمند بازنگری است. تمرکز باید بر «قابلیت پیش‌بینی قابل قبول» سامانه‌ها باشد، به‌گونه‌ای که عملکرد آن‌ها با استانداردهای IHL منطبق و قابل ارزیابی باشد. این چارچوب، مسیر طراحی، آزمون و تأیید سامانه‌ها را روشن کرده و نظارت انسانی را در مرحله‌ای اعمال می‌کند که بیشترین اثرگذاری را دارد.

در نهایت، با طراحی و بهره‌گیری مسئولانه از سامانه‌های یادگیرنده، می‌توان ضمن رعایت اصول اساسی IHL، از جمله ضرورت نظامی و کاهش خسارات غیرضروری، از ظرفیت‌های فناورانه برای کاهش تلفات غیرنظامیان بهره برد. رویکرد پیشنهادی، تعادل میان قابلیت‌های فناورانه و مسئولیت انسانی را برقرار ساخته و امکان استفاده بهینه از فناوری‌های نوین در چارچوب حقوق بین‌الملل بشردوستانه را فراهم می‌آورد.

واژگان کلیدی: حقوق بین‌الملل بشردوستانه، سلاح‌های خودکار، کنترل انسانی، هوش مصنوعی، یادگیری ماشینی، پیش‌بینی‌پذیری.

سرآغاز

به‌کارگیری فناوری‌های پیشرفته هوش مصنوعی و یادگیری ماشینی در حوزه نظامی، در حال دگرگونی ماهوی جنگ، به‌ویژه در فرآیند تصمیم‌گیری برای استفاده از نیروی کُشنده است (de Spiegeleire, Maas & Sweijs, 2017: 35–39). سامانه‌های تسلیحاتی مبتنی بر این فناوری‌ها که قادر به یادگیری از محیط و تطبیق رفتار خود هستند، اکنون در کانون یک رقابت تسلیحاتی نوین میان قدرت‌های بزرگ قرار گرفته‌اند (The Economist, 2018; Lant, 2017). برخی از صاحبنظران، ظهور این فناوری در عرصه جنگ را با انقلابی که اختراع سلاح هسته‌ای ایجاد کرد، قابل قیاس دانسته‌اند (Simonite, 2017). این تحول، اصل بنیادین کنترل انسانی بر عملیات نظامی را که ریشه در حقوق بین‌الملل بشردوستانه[1] دارد، با پرسشی بی‌سابقه مواجه ساخته است (International Committee of the Red Cross, 2014: 7).

ویژگی ذاتی سامانه‌های یادگیرنده، یعنی قابلیت «یادگیری» مستمر از محیط و بازخوردهای عملیاتی، سبب می‌شود رفتار و فرآیند تصمیم‌گیری آن‌ها برای انسان غیرقابل‌پیش‌بینی و غیرقابل فهم باشد (Matthias, 2004: 175). این «غیر قابل پیش‌بینی بودن» و «عدم شفافیت» (مشهور به مسئله «جعبه سیاه») (Mittelstadt et al., 2016: 11; Knight, 2017)، اساساً امکان اعمال آن شکل از نظارت و مداخله مستقیم انسانی را که در الگوهای سنتی «کنترل انسانی معنادار» متصور است، منتفی می‌سازد (Roff & Moyes, 2016). بر این مبنا، استدلال قوی‌ای مطرح شده که چنین سامانه‌هایی، به دلیل عدم امکان قرار گرفتن تحت کنترل معنادار انسان، ذاتاً با IHL ناسازگار و درنتیجه باید ممنوع شوند (Grut, 2013: 5; Human Rights Watch, 2015).

---

1. International Humanitarian Law (IHL)

با این حال، تمرکز انحصاری ادبیات موجود بر چالش «فقدان کنترل عملیاتی»، سبب غفلت از پرسشی بنیادی‌تر شده است (Deeks, Lubell & Murray, 2019; Matthias, 2004; Schuller, 2017). پرسش اصلی این است: آیا برای تضمین رعایت مقررات IHL توسط یک سامانه خودکار یادگیرنده، همواره و در همه شرایط، وجود «کنترل انسانی» به شکل سنتی آن (مداخله در حلقه عملیاتی) ضروری است؟ به بیان دیگر، آیا ممکن است توانمندی‌های تحلیلی منحصربه‌فرد این سامانه‌ها (مانند پردازش حجم عظیم داده و محاسبات پیچیده) در برخی سناریوها، منجر به تصمیم‌گیری منطبق‌تر با اصول تفکیک و تناسب در مقایسه با یک اپراتور انسانی تحت فشار شود؟ (Schuller, 2017: 411; Silver & Hassabis, 2017). اگر پاسخ مثبت باشد، الزام جزمی به الگوی خاص کنترل لحظه‌ای ممکن است نه‌تنها کمکی به حمایت از غیرنظامیان نکند، بلکه با محروم کردن طرفین درگیری از یک ابزار بالقوه کاهنده تلفات، نتیجه‌ای عکس دهد.

این مقاله در پی پاسخ به این پرسش و باز تعریف مفهوم کنترل انسانی در عصر سلاح‌های هوشمند است. فرضیه مقاله آن است که رعایت IHL را می‌توان محقق ساخت. برای آزمون این فرضیه، پژوهش حاضر با تحلیل محتوای قواعد اولیه و ثانویه IHL (مانند اصل تفکیک و تعهد به مراقبت مستمر) و نیز نقد و ارزیابی دکترین‌های مطرح در زمینه کنترل انسانی، استدلال خود را پیش می‌برد. بر این اساس مقاله در بخش نخست، به مروری فشرده بر مبانی فنی سامانه‌های یادگیرنده می‌پردازد. بخش دوم، جایگاه این سامانه‌ها را در چارچوب IHL، به‌ویژه در پرتو «تعهد به مراقبت مستمر»[1] (Protocol Additional to the

1. Constant Care Obligation: ماده ۵۷ پروتکل اول الحاقی کنوانسیون‌های ژنو؛ به‌موجب این تعهد هر اقدامی در جنگ باید با مراقبت دائمی و هوشیاری مداوم برای حفظ جان غیرنظامیان همراه باشد.

(Geneva Conventions, 1977: art. 57)، تحلیل کرده و کاستی‌های دکترین‌های کنترلی موجود را بررسی می‌کند. درنهایت بخش سوم، چارچوب پیشنهادی مقاله برای بازتعریف کنترل انسانی را ارائه خواهد داد.

1. مروری بر مبانی فنی سامانه‌های تسلیحاتی خودکار

درک ماهیت و قابلیت‌های سامانه‌های تسلیحاتی خودکار مبتنی بر هوش مصنوعی، پیش‌نیاز ضروری برای تحلیل دقیق چالش‌های حقوقی آن‌هاست. این بخش به‌طور فشرده، دو جنبه کلیدی فناوری را که مستقیماً مباحث حقوق بشردوستانه را تحت تأثیر قرار می‌دهند، بررسی می‌کند؛ نخست، الگوهای تعامل انسان و ماشین که چارچوب بحث «کنترل» را مشخص می‌کنند؛ و دوم، نقش فناوری یادگیری ماشینی که با تغییر ماهیت خودکارسازی، بنیان‌های سنتی کنترل و مسئولیت‌پذیری را متحول ساخته است.

1-1. تعریف و ساختار سامانه‌های تسلیحاتی خودکار: چارچوبی برای تحلیل چالش کنترل

در حوزه حقوقی و فناورانه، تعریف دقیق و مورد اجماع از سامانه‌های تسلیحاتی خودکار (AWS) هنوز ارائه نشده است (INT'L COMM. OF THE RED CROSS, 2014: 7). با این حال، تعاریف موجود عموماً بر دو ویژگی کلیدی که مستقیماً به مباحث حقوق بشردوستانه مرتبط می‌شوند تأکید دارند: «خودکارسازی استفاده از نیروی کُشنده» و «جایگزینی نیروی انسانی در چرخه هدفگیری» (INT'L COMM. OF THE RED CROSS, 2014: 7; Rep. of Switzerland, 2016: 1). این دو ویژگی، هسته چالشی را تشکیل می‌دهند که این مقاله به آن می‌پردازد: چگونه می‌توان در فقدان کنترل مستقیم و سنتی انسانی،

اطمینان حاصل کرد که «وظایف محوله تحت حاکمیت حقوق بین‌الملل بشردوستانه (IHL)» انجام می‌شوند؟

1-1-1. طیف تعامل انسان و ماشین: از نظارت تا خودمختاری و پیامد آن برای کنترل

سامانه‌های خودکار را می‌توان بر اساس طیف تعامل بین انسان و ماشین به شرح زیر دسته‌بندی کرد:

- **تسلیحات فرمان‌بر:** سامانه‌ها تنها از طریق فرمان انسان هدف‌ها را انتخاب و نیرو را اعمال می‌کنند.
- **تسلیحات تحت نظارت: انتخاب** هدف و اعمال نیرو توسط سامانه اما تحت نظارت انسان است که می‌تواند در صورت نیاز تصمیم سامانه را لغو یا اصلاح کند.
- **تسلیحات مستقل:** انتخاب هدف و اعمال نیرو مستقلاً توسط سامانه‌ها بدون دخالت یا کنترل انسان است (Human Rights Watch, 2015).

نکته حیاتی برای بحث حاضر این است که حتی در مدل «تحت نظارت» نیز ماهیت فناوری‌های پیشرفته یادگیری ماشینی - که موضوع اصلی این مقاله است - می‌تواند کیفیت و امکان واقعی این نظارت انسانی را به‌طور بنیادین زیر سؤال ببرد؛ بنابراین، بررسی این دسته‌بندی‌ها مقدمه‌ای است برای پرداختن به این پرسش که آیا «نظارت» یا «کنترل» به شکل سنتی، معیار مناسبی برای ارزیابی این سامانه‌های نوین است؟

1-1-2. خودکارسازی فرآیند تصمیم‌گیری: انتقال از عمل به قضاوت

تحلیل دقیق‌تر این چالش، مستلزم توجه به مدل چرخه تصمیم‌گیری اودا[1] (مشاهده[2]، جهت‌یابی[3]، تصمیم‌گیری[4] و اقدام[5]) است (Marra & McNeil, 2013: 1145). در این مدل، خودکارسازی تاریخی عمدتاً در مرحله «مشاهده» (مانند استفاده از حسگرها) متوقف شده بود که IHL با آن مشکلی ندارد (Boulanin & Verbruggen, 2017: 27–29؛ Schuller, 2017: 394). آنچه فناوری یادگیری ماشینی را به یک مسئله حقوقی بی‌سابقه تبدیل می‌کند، امکان خودکارسازی مراحل «جهت‌یابی» و «تصمیم‌گیری» است. در این مراحل، سامانه به‌طور مستقل داده‌ها را تحلیل، گزینه‌های عملی را ارزیابی و مسیر «بهینه» اقدام را - که می‌تواند استفاده از نیروی کُشنده باشد - انتخاب می‌کند (Schuller, 2017: 394, 396-97).

این تحول به معنای واگذاری قضاوت‌های مبتنی بر ارزش و ملاحظات پیچیده انسانی (نظامی و بشردوستانه) به ماشین است. همان‌طور که شولر اشاره می‌کند، این امر پیوند علی مستقیم بین قضاوت انسانی و نتیجه مرگبار را تضعیف می‌کند (Schuller, 2017: 394-97)؛ بنابراین، پرسش محوری این

---

1. OODA Loop؛ چرخه تصمیم‌گیری نظامی متشکل از چهار مرحله «مشاهده (Observe)»، جهت‌یابی (Orient)، تصمیم‌گیری (Decide) و اقدام (Act)» که نخستین بار توسط جان بوید مطرح شد و برای تحلیل فرایند تصمیم‌گیری انسان و ماشین در شرایط درگیری نظامی به کار می‌رود.
2. Observe (مشاهده)؛ مرحله نخست چرخه OODA که شامل جمع‌آوری داده‌ها و اطلاعات از محیط عملیاتی، ازجمله داده‌های حسی، اطلاعات میدانی و ورودی‌های محیطی است.
3. Orient (جهت‌یابی / تفسیر)؛ مرحله‌ای در چرخه OODA که در آن اطلاعات خام دریافتی تحلیل، تفسیر و در چارچوب الگوها، تجربیات پیشین، قواعد و اهداف عملیاتی معنا گذاری می‌شود.
4. Decide (تصمیم‌گیری)؛ مرحله‌ای از چرخه OODA که در آن، بر اساس تفسیر اطلاعات و گزینه‌های موجود، یک مسیر اقدام انتخاب می‌شود.
5. Act (اقدام)؛ مرحله نهایی چرخه OODA که شامل اجرای تصمیم اتخاذ شده در محیط عملیاتی است و نتایج آن مجدداً وارد مرحله مشاهده می‌شود.

مقاله نه صرفاً درباره خودکارسازی یک «عمل» فیزیکی که درباره خودکارسازی «فرآیند قضاوت و تصمیم‌گیری» است که در قلب قواعد IHL قرار دارد. این انتقال، نیازمند بازتعریف مفاهیمی چون کنترل، مسئولیت و مراقبت است که در بخش‌های بعدی مورد کاوش قرار می‌گیرد.

2-1. یادگیری ماشینی و نقش آن در سامانه‌های تسلیحاتی خودکار

سامانه‌های تسلیحاتی خودکار آینده، به‌جای تکیه‌بر قواعد ساده شرطی، بر فناوری یادگیری ماشینی متکی خواهند بود. این گذار از «خودکارسازی» به «یادگیری»، پیامدهای بنیادینی برای حقوق بشردوستانه دارد؛ زیرا هسته کنترل را از اجرای دستورالعمل‌های ثابت به توانایی تطبیق پویا و یادگیری از تجربه تغییر می‌دهد (Boulanin & Verbruggen, 2017: 16-17). این بخش، دو ویژگی کلیدی این فناوری را که چالش حقوقی ایجاد می‌کنند - یادگیری عمیق و یادگیری تقویتی - بررسی کرده و سپس پیامدهای حقوقی ناشی از ترکیب آن‌ها را برمی‌شمارد.

۱-۲-۱. یادگیری عمیق: قدرت تشخیص و مسئله «جعبه سیاه»

یادگیری عمیق با الهام از شبکه‌های عصبی مغز، به سیستم‌ها امکان می‌دهد تا مستقیماً از داده‌های خام الگوهای پیچیده را استخراج کنند (Boulanin & Verbruggen, 2017: 17). این توانایی، آن را برای وظایفی مانند شناسایی و تشخیص هدف در داده‌های حجیم اطلاعاتی (مانند تصاویر ماهواره‌ای یا سیگنال‌های راداری) ایده آل می‌سازد؛ زیرا می‌تواند تمایزهایی را انجام دهد که حتی برنامه‌نویس انسانی قادر به کدگذاری آن نیست (Mittelstadt et al., 2016: 6). با این حال همین قدرت، بزرگترین نقطه‌ضعف حقوقی آن است: فرآیند

تصمیم‌گیری در شبکه‌های عصبی عمیق، برای انسان غیرقابل فهم (غیر شفاف) است؛ پدیده‌ای که به «جعبه سیاه» شهرت دارد (Matthias, 2004: 179). همان‌گونه که در تحلیل قراردادهای الگوریتمی جعبه سیاه اشاره شده است، پیش‌بینی رفتار و تبیین فرآیند تصمیم‌گیری این سامانه‌ها حتی برای طراحان آن‌ها نیز دشوار است (علائی و حسین زاده، ۱۴۰۱: ۲۵۴). این عدم شفافیت، امکان نظارت معنادار انسانی بر تصمیمات حیاتی را به‌شدت تضعیف می‌کند.

2-۲-1. یادگیری تقویتی: بهینه‌سازی خودمختار و افزایش عدم قطعیت

یادگیری تقویتی گامی فراتر می‌نهد. در این روش، سیستم نه از داده‌های ایستا، بلکه از طریق تعامل مستقیم و آزمون‌وخطا با محیط آموزش می‌بیند و رفتار خود را برای حداکثر کردن یک «پاداش» از پیش تعریف شده (مانند دستیابی به یک هدف نظامی) بهینه می‌کند (Boulanin & Verbruggen, 2017: 17; Silver et al., 2018). نمونه بارز آن، سیستم AlphaGo Zero است که بدون هیچ دانش اولیه انسانی، تنها با بازی مقابل خود، به استراتژی‌هایی خلاقانه و فرا بشری دست یافت (Silver & Hassabis, 2017; Kania, 2017). این قابلیت، سامانه را برای محیط‌های پیچیده و غیرقابل‌پیش‌بینی میدان نبرد مناسب می‌سازد؛ اما هم‌نشینی دو ویژگی – یادگیری از محیط و بهینه‌سازی برای یک هدف – منجر به غیرقابل‌پیش‌بینی بودن ذاتی می‌شود (Additional Protocol I, 1977: art. 57). سیستم ممکن است راه‌حل‌هایی بیابد که برنامه‌ریزان انسانی هرگز پیش‌بینی نکرده‌اند (Mittelstadt et al., 2016: 3–4) که این امر مسئولیت‌پذیری و ارزیابی پیشینی انطباق با قانون را دشوار می‌سازد.

3-۲-1. ابعاد حقوقی پیچیده: از تعهد مراقبت تا سوگیری داده‌ها

ترکیب قابلیت‌های فوق، سه چالش حقوقی متمایز ولی مرتبط ایجاد می‌کند:

تعارض با تعهد به مراقبت مستمر: فرآیند تصمیم‌گیری غیرقابل فهم (جعبه سیاه یادگیری عمیق) و غیرقابل‌پیش‌بینی (یادگیری تقویتی)، امکان اعمال «تعهد پیشگیرانه مراقبت مستمر» را که برای حفاظت از غیرنظامیان ضروری است، عملاً ناممکن می‌سازد (Knight, 2017). چگونه می‌توان بر سیستمی مراقبت کرد که نحوه تصمیم‌گیری‌اش را درک نمی‌کنیم و رفتار آینده‌اش را نمی‌توانیم پیش‌بینی کنیم؟

معضل مسئولیت کیفری: از آنجا که پیش‌بینی رفتار نهایی این سامانه‌ها حتی برای توسعه‌دهندگانش نیز دشوار است (Bishop, 2006)، احراز عنصر روانی لازم برای مسئولیت فردی در حقوق کیفری بین‌المللی (مثلاً قصد یا علم به نقض قانون) در صورت بروز حادثه، با مشکل مواجه می‌شود (Kania, 2017). این چالش منحصر به عرصه نظامی نیست و ناکارآمدی قوانین سنتی در تخصیص مسئولیت حوادث ناشی از سامانه‌های خودمختار در حوزه‌های غیرنظامی نیز مورد تأکید قرار گرفته است (پارسا، ۱۴۰۳: ۸۷). این «شکاف مسئولیت» می‌تواند به مصونیت عملی از مجازات بینجامد.

چالش نوین: تهدید سوگیری در داده‌های آموزشی (حفره تحلیل پیشین). نکته‌ای که در ادبیات فعلی اغلب نادیده گرفته می‌شود، این است که عملکرد و قضاوت یک سامانه یادگیرنده، بازتاب مستقیم داده‌هایی است که با آن آموزش دیده است. اگر داده‌های آموزشی ناقص، تاریخ گذشته یا حاوی سوگیری‌های نظامی، فرهنگی یا جمعیتی باشند، سامانه این سوگیری‌ها را در عملیات واقعی تقویت و اجرا خواهد کرد. این امر می‌تواند به نقض سیستماتیک اصول تفکیک و تناسب

منجر شود، حتی اگر الگوریتم از نظر فنی «درست» کار کند. مدیریت این ریسک، نیازمند نظارت و ممیزی دقیق در مرحله‌ای است که در مدل‌های کنونی کنترل انسانی نادیده گرفته می‌شود: مرحله آموزش و تأیید پیش از به‌کارگیری.

با این حال، این فناوری توانمندی بی‌بدیلی نیز دارد: پردازش حجم عظیم داده با سرعت و دقت فرا بشری. این توانایی، در تئوری می‌تواند منجر به تصمیمات دقیق‌تر، شناسایی بهتر اهداف مشروع و کاهش خسارت جانبی نسبت به یک فرمانده انسانی تحت فشار شود؛ بنابراین پرسش اساسی این نیست که آیا این فناوری خطرناک است، بلکه این است: آیا می‌توان چارچوبی حقوقی طراحی کرد که هم از این فرصت بهره ببرد و هم بر این سه چالش (شفافیت، پیش‌بینی پذیری و سوگیری داده) فائق آید؟ پاسخ این مقاله به این پرسش، در ادامه و با بازتعریف مفهوم کنترل انسانی ارائه خواهد شد.

3-1. کاربردهای کنونی و افق آینده یادگیری ماشینی در حوزه نظامی

یادگیری ماشینی هم‌اکنون نیز در حوزه نظامی حضوری فعال دارد؛ اما این حضور عمدتاً در چارچوب افزایش دقت و کارایی سیستم‌های موجود و در نقش ابزار کمکی برای تصمیم‌گیرنده انسانی تعریف شده است. نمونه بارز آن، استفاده از یادگیری عمیق در سامانه‌های خودکار تشخیص هدف (ATR) است که با تحلیل داده‌های حسگرها، دقت شناسایی اهداف را افزایش و نرخ هشدارهای اشتباه را کاهش می‌دهد (Host, 2016). این فناوری به سیستم کمک می‌کند تا در محیط‌های پیچیده، بهتر در برابر اقدامات فریبنده مقاومت کند و خطاهای ادراکی را به حداقل برساند (de Spiegeleire, Maas & Sweijs, 2017: 88–89).

با این حال، گذر از این کاربردهای کمکی به خودکارسازی کامل چرخه هدفگیری با موانع فنی عمده‌ای روبروست. طراحی سامانه‌ای که بتواند تمامی حالات ممکن در یک محیط عملیاتی ـ حتی نسبتاً پایدار ـ را پوشش دهد، مستلزم دسته داده‌های عظیم و تقریباً کامل است که تهیه آن در عمل بسیار دشوار یا غیرممکن می‌کند (Schuller, 2017: 410; see also Boulanin & Verbruggen, 2017: 65–82). به همین دلیل، در وضعیت کنونی، سیستم‌های مبتنی بر یادگیری عمیق در نقش تأیید کننده یا توصیه گر برای اپراتور انسانی عمل می‌کنند و تصمیم نهایی و مسئولیت آن بر عهده انسان باقی می‌ماند (Boulanin & Verbruggen, 2017: 25–26).

اما تأکید این مقاله بر افق آینده و قابلیت‌های بالقوه این فناوری است. با توجه به شتاب تحولات فنی و سرمایه‌گذاری سنگین قدرت‌های نظامی در این عرصه، بررسی آثار حقوقی خودکارسازی تصمیم‌گیری در چرخه هدفگیری - حتی اگر امروز محقق نشده باشد - یک ضرورت پیشگیرانه و آینده‌نگرانه است. تمرکز بر محدودیت‌های فنی کنونی نباید باعث غفلت از تحلیل این پرسش کلیدی شود که: اگر روزی موانع فنی مرتفع شدند و یک سامانه یادگیرنده بتواند مستقل از انسان، تصمیم به استفاده از نیروی کُشنده بگیرد، آیا چارچوب حقوقی فعلی برای ارزیابی مشروعیت آن کافی است؟ این پرسش، موضوع بخش‌های تحلیلی بعدی این مقاله را تشکیل می‌دهد.

2. جایگاه سامانه‌های تسلیحاتی خودکار مبتنی بر یادگیری ماشینی در حقوق بین‌الملل بشردوستانه

سامانه‌های تسلیحاتی خودکار مبتنی بر یادگیری ماشینی، در تقاطع دو حوزه پیچیده فناوری پیشرفته و حقوق بشردوستانه قرار گرفته‌اند.

حقوق بین‌الملل بشردوستانه (IHL) به‌عنوان چارچوبی پایدار اما انعطاف‌پذیر، متشکل از اصول کلی و قواعد خاص است که در مواجهه با فناوری‌های نوین جنگ مورد تفسیر و اعمال قرار می‌گیرد. پرسش محوری این نیست که آیا IHL بر این سامانه‌ها حاکم است (که پاسخ آن قطعاً مثبت است)، بلکه این است که آیا اصول و قواعد موجود IHL - که عمدتاً بر پایه عامل انسانی و قضاوت انسانی بنا شده‌اند - برای تنظیم رفتار سامانه‌هایی که ذاتاً فاقد قضاوت انسانی و درعین‌حال غیرقابل‌پیش‌بینی هستند، کافی و کارآمد است؟ این بخش از مقاله با تحلیل دو محور کلیدی به این پرسش می‌پردازد: نخست، بررسی چالش بنیادین غیرقابل‌پیش‌بینی بودن و پیامدهای آن برای مسئولیت و الزامات احتیاطی؛ و دوم، واکاوی چگونگی اعمال اصل اساسی «تعهد به مراقبت مستمر» بر این سامانه‌ها و نقش کنترل انسانی در تحقق آن. درنهایت، این تحلیل مبنایی برای نقد رویکردهای موجود و ارائه چارچوب جایگزین در بخش‌های آتی فراهم می‌کند.

2-1. غیرقابل‌پیش‌بینی بودن ذاتی و چالش‌های حقوقی آن در چارچوب IHL

پیش‌بینی‌پذیری به‌عنوان یک شرط بنیادین برای اعمال مسئولیت در حقوق بین‌الملل بشردوستانه (IHL) و حقوق کیفری بین‌المللی (ICL) عمل می‌کند. در حالی که سامانه‌های خودکار سنتی مبتنی بر قواعد شرطی، به دلیل تعیین دقیق خروجی برای هر ورودی توسط برنامه‌نویس، ماهیتی قابل پیش‌بینی دارند (Scharre, 2016)، ماهیت سامانه‌های تسلیحاتی خودکار یادگیرنده‌ای[1] این اصل را در معرض تهدید قرار می‌دهد. این

1. «Learning AWS»: به سامانه‌های تسلیحاتی خودکار گفته می‌شود که قابلیت یادگیری دارند و می‌توانند بر اساس تجربه یا داده‌های جدید، رفتار خود را اصلاح و بهبود دهند.

سامانه‌ها به دلیل قابلیت «یادگیری» و انطباق مستمر با محیط، ذاتاً غیرقابل‌پیش‌بینی هستند (Int'l Comm. of the Red Cross, 2016). این غیرقابل‌پیش‌بینی بودن، دو چالش حقوقی عمده ایجاد می‌کند:

- **معضل مسئولیت کیفری:** بر اساس اصول ICL،[1] مسئولیت فردی برای نقض‌های جدی IHL تنها نسبت به پیامدهایی قابل انتساب است که «به‌طور معقول قابل پیش‌بینی» باشند (Horowitz, Scharre & Center for a New American Security, 2015). حال اگر یک سامانه یادگیرنده که پیش‌تر مطابق با IHL عمل می‌کرده است، به دلیل یک الگوی یادگیری غیرمنتظره یا واکنش به یک محرک محیطی پیش‌بینی نشده، رفتاری انجام دهد که منجر به آسیب به غیرنظامیان شود، احراز این عنصر «پیش‌بینی معقول» برای انتساب مسئولیت به هر انسان مرتبط (طراح، فرمانده یا اپراتور) دشوار یا ناممکن خواهد بود. این خلأ، می‌تواند منجر به یک «شکاف مسئولیت» شود که در آن نقض آشکار IHL بدون یافتن مقصر انسانی مشخص، باقی می‌ماند.[2]
- **مانع اجرای الزامات احتیاطی:** چالش فراتر از مسئولیت پسینی است. اصل «تعهد به مراقبت

---

1. «International Criminal Law» یا ICL: شاخه‌ای از حقوق بین‌الملل است که به تعقیب و مجازات شدیدترین جرائم بین‌المللی ازجمله جنایات جنگی، جنایات علیه بشریت و نسل‌کشی می‌پردازد. همه نقض‌های IHL به‌صورت خودکار جرائم بین‌المللی محسوب نمی‌شوند. «حقوق کیفری بین‌المللی (ICL) تنها یکی از ابزارهای اجرای حقوق بین‌الملل بشردوستانه (IHL) است. عموماً تعقیب کیفری تحت ICL محدود به «جدی‌ترین جرائم موردتوجه جامعه بین‌المللی» می‌شود. مطابق مقدمه و ماده ۲۵ (۲) اساسنامه رم دیوان کیفری بین‌المللی (۱۷ ژوئیه ۱۹۹۸، U.N.T.S. 90 2187؛ لازم‌الاجرا از ۱ ژوئیه ۲۰۰۲)، بنابراین هرگونه نقض IHL به‌طور خودکار جرم بین‌المللی محسوب نمی‌شود.»

2. «برای مثال، ماده ۳۰ همان مرجع بیان می‌کند: ۱) برای احراز یک جرم بر اساس اساسنامه رم، عنصر روانی لازم (mens rea) باید وجود داشته باشد؛ و ۲) استاندارد عمومی عنصر روانی که پایین‌تر از قصد (Intent) است، عبارت است از «دانش» یا «آگاهی»، یعنی «آگاهی از وجود یک وضعیت یا از وقوع نتیجه‌ای در جریان عادی امور».»

مستمر» و سایر الزامات احتیاطی IHL مستلزم آن است که فرماندهان اقدامات عملی برای جلوگیری از نقض‌ها انجام دهند؛ اما چگونه می‌توان بر سامانه‌ای که رفتارش ذاتاً غیرقابل‌پیش‌بینی است «مراقبت» کرد؟ این پرسش اساسی، محور بحث این مقاله را تشکیل می‌دهد: با فرض غیرقابل‌پیش‌بینی بودن ذاتی سامانه‌های یادگیرنده، آیا الزامات احتیاطی IHL همچنان ایجاب می‌کند که حداقلی از «نظارت انسانی» در حین عملیات بر آن‌ها اعمال شود؟ یا اینکه پارادایم کنترل باید به‌گونه‌ای بنیادین بازتعریف شود تا تمرکز از «نظارت در حین عمل» به «تضمین کیفیت و اعتبارسنجی پیش از عمل» منتقل شود؟

غلبه نگاه سنتی در دکترین حقوقی بر مبنای دیدگاه اول (لزوم نظارت عملیاتی) استوار است. با این حال، این مقاله با پذیرش واقعیت غیرقابل‌پیش‌بینی بودن، در پی بررسی امکان‌پذیری و برتری دیدگاه دوم (انتقال کانون کنترل) در چارچوب IHL است.

۲-2. الزامات احتیاطی در حقوق بین‌الملل بشردوستانه: اصول بنیادین و ابزارهای عملیاتی

حقوق بین‌الملل بشردوستانه (IHL) نه مجموعه‌ای از قواعد صرفاً ممنوعه، بلکه چارچوبی متوازن کننده است که می‌کوشد میان ضرورت نظامی مشروع و ملاحظات انسانی تعادل برقرار سازد (Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996: paras. 261, 493). این تعادل از طریق اصول بنیادینی که زیرساخت قواعد خاص معاهداتی و عرفی را تشکیل می‌دهند عملیاتی می‌شود.

دو اصل ساختاری که مستقیماً بر عملیات نظامی حاکم‌اند، اصل تفکیک[1] و اصل تناسب[1] هستند. اصل

1. Distinction با اصل تمایز: تمایز بین رزمندگان/نیروهای نظامی و غیرنظامیان در هدفگیری و حمله، به‌گونه‌ای که حمله صرفاً بر اهداف نظامی مجاز باشد.

تفکیک، بنیان حمله مشروع را مشخص می‌سازد و الزام می‌کند که حملات صرفاً متوجه اهداف نظامی باشد و از افراد غیرنظامی و اموال غیرنظامی حمایتی مطلق به عمل آورد. اصل تناسب که مکمل اصل تفکیک است، اجازه می‌دهد در حمله به یک هدف نظامی مشروع، آسیب جانبی به غیرنظامیان و اموال غیرنظامی وارد آید، اما این آسیب نباید با منفعت نظامی ملموس و مستقیم مورد انتظار از حمله نامتناسب باشد.[2] این اصول، در کنار اصل انسانیت و اصل ضرورت نظامی (Hague Convention (IV) Respecting the Laws and Customs of War on Land, preamble: 26 Stat. 2277)، چهارچوب هنجاری کلی IHL را تعریف می‌کنند.

با این حال، این اصول کلی برای هدایت رفتار در میدان نبرد نیاز به ابزارهای عملیاتی دارند.[3] این ابزارها در قالب الزامات احتیاطی تجلی یافته‌اند. الزامات احتیاطی، مجموعه‌ای از اقدامات فعال و پیشگیرانه‌اند که طرف‌های درگیر مخاصمه ملزم به رعایت آن‌ها در حین برنامه‌ریزی، تصمیم‌گیری و اجرای عملیات هستند تا اطمینان حاصل شود که اصول تفکیک و تناسب در عمل رعایت

---

1 Proportionality یا اصل تناسب: در حملات نظامی، میزان خسارت جانبی به غیرنظامیان نباید نامتناسب با مزیت نظامی پیش‌بینی‌شده باشد.

2. برای نمونه، اصل تمایز همان‌طور که در مواد ۴۸ و ۵۲ پروتکل الحاقی اول کنوانسیون‌های ژنو مقرر شده است، تعیین می‌کند چه کسی رزمنده است و چه اهداف نظامی می‌توانند به‌طور مجاز مورد حمله قرار گیرند؛ همچنین، اصل تناسب در مواد ۵۱ (۵) (ب)، ۵۷ (۲) (الف) (iii) و ۵۷ (۲) (ب) پروتکل الحاقی اول منعکس شده است و ضوابط ارزیابی تناسب خسارات جانبی و ضرورت حمله را مشخص می‌کند.

3. Precautionary obligations یا تعهدات احتیاطی: اقداماتی که برای کاهش آسیب جانبی به غیرنظامیان و اموال غیرنظامی باید در طراحی و اجرای عملیات نظامی انجام شود (ماده 57 پروتکل اول الحاقی).

می‌شوند.[1] هسته مرکزی این الزامات، تعهد به مراقبت مستمر برای حفظ جمعیت غیرنظامی و اهداف غیرنظامی است که در ماده ۵۷ پروتکل الحاقی اول به کنوانسیون‌های ژنو تصریح شده است (Additional Protocol I, 1977: art. 57) و در حقوق عرفی نیز به رسمیت شناخته شده است (Henckaerts & Doswald-Beck, 2005: 51; Pilloud et al., 1987: 2191).

در مواجهه با سامانه‌های تسلیحاتی خودکار یادگیرنده، پرسش محوری این است که محتوای عملی این تعهد به مراقبت مستمر چیست؟ آیا این تعهد، مستلزم حفظ کنترل انسانی سنتی در حلقه عملیاتی است یا می‌توان آن را از طریق تدابیر احتیاطی جایگزین در مراحل طراحی، آزمایش و استقرار سامانه محقق ساخت؟ پاسخ به این پرسش، مستلزم تحلیل دقیق مفهوم «مراقبت» و نسبت آن با ماهیت غیرقابل‌پیش‌بینی و خودمختار این فناوری نوین است.

2-3. سامانه‌های تسلیحاتی خودکار و تکلیف مراقبت دائمی: ابهام مفهومی و چالش تفسیر

تکلیف «مراقبت دائمی» مندرج در ماده ۵۷ (۱) پروتکل الحاقی اول (Additional Protocol I: art. 57(1))، به‌عنوان هسته مرکزی و عامل توانمند ساز سایر الزامات احتیاطی عمل می‌کند و تضمین می‌کند که اصول تفکیک و تناسب در عمل محقق شوند (Protective Edge, 2015). با وجود اهمیت بنیادین، محتوای عینی این تکلیف در هاله‌ای از ابهام باقی مانده است. این اصطلاح در متون حقوقی تعریف نشده و حتی تفسیر معتبر کمیته بین‌المللی صلیب سرخ نیز صرفاً از آن به‌عنوان یک «اصل کلی» یاد می‌کند (Commentary on the Additional Protocols, 1987: 2191). درحالی‌که دامنه شمول آن به تمامی حوزه‌ها و

1. ماده ۵۷ پروتکل اول الحاقی کنوانسیون‌های ژنو: به موجب این ماده، هر اقدام در جنگ باید با مراقبت مداوم و هوشیاری برای حفظ جان غیرنظامیان همراه باشد.

سطوح عملیات جنگی گسترده است (Program on Humanitarian Policy and Conflict Resolution, 2010: 124–125)، استاندارد دقیق «میزان مراقبت» الزامی مشخص نشده است (Gill et al., 2014: 15).

رویکرد قضایی و معیار «عملی بودن»: کمیته حقیقت‌یاب دیوان کیفری بین‌المللی برای یوگسلاوی سابق در پرونده عملیات ناتو،[1] در تفسیر این تکلیف، معیار تعیین‌کننده‌ای را ارائه کرد.[2] این نهاد تأکید کرد که الزام به مراقبت دائمی، تنها در چارچوب آنچه در شرایط عملیاتی مشخص «مقدور و عملی» است معنا می‌یابد و به فرماندهان در انتخاب ابزار و روش‌های تحقق این تکلیف، «دامنه‌ای از اختیارات» اعطا می‌کند (Int'l Crim. Trib. for the Former Yugoslavia, 2000: 29, 1257)؛ بنابراین، استاندارد مراقبت دائمی بر پایه ارزیابی آنچه برای یک فرمانده در شرایط خاص «عملی» محسوب می‌شود استوار گشته است (Schmitt, 2013: 20; Wagner, 2014: 1397). همین ویژگی، ماهیت ذاتی قضاوتی و مبتنی بر شرایط به آن بخشیده و تعریف دقیق و جهان‌شمول آن را دشوار ساخته است.

تشدید ابهام در مواجهه با فناوری‌های نوین: این ابهام ذاتی، با ظهور فناوری‌های پیچیده‌ای مانند سامانه‌های تسلیحاتی خودکار یادگیرنده که فاقد پیشینه قضایی یا مقرره گذاری مشخص هستند، به اوج می‌رسد. تلاش‌های انجام‌شده برای تطبیق اصول حقوقی با فناوری‌های نوین - مانند «راهنمای تالین درباره حقوق بین‌الملل ناظر بر

---

1. ICTY: مخفف International Criminal Tribunal for the former Yugoslavia (دادگاه جنایی بین‌المللی یوگسلاوی سابق) است. این دادگاه در دهه ۱۹۹۰ برای رسیدگی به جنایات جنگی، جنایات علیه بشریت و نسل‌کشی در جنگ‌های بالکان تشکیل شد و هدفش محاکمه مسئولان این جرائم بود.
2. این گزارش نظر کارشناسی ارائه می‌دهد در مورد معنای «احتیاطات قابل انجام» (Feasible precautions) در حقوق بین‌الملل بشردوستانه و ارزیابی اقدامات طرفین در درگیری غزه در چارچوب عملیات «لبه حفاظتی».

عملیات سایبری» (Schmitt, 2013, 2017) - گرچه مفید هستند، اما ابهامات بنیادین را مرتفع نمی‌سازند. این راهنما، هرچند تأکید می‌کند که تکلیف مراقبت دائمی کلیه مراحل عملیات از برنامه‌ریزی تا اجرا را در برمی‌گیرد و مستلزم حساسیت مستمر فرماندهان نسبت به پیامدهای اقداماتشان است (Schmitt, 2017:1-12, 477)، اما درنهایت تصریح می‌کند که در مواجهه با پیچیدگی‌های فنی، «برنامه‌ریزان عملیات باید، در صورت امکان، از متخصصان فنی بهره بگیرند» (Schmitt, 2017: 477). این رهنمود، اگرچه به «سامانه‌های یادگیرنده» نیز قابل‌تعمیم است، پرسش محوری را بی‌پاسخ می‌گذارد: آیا تکلیف مراقبت دائمی، در مورد این سامانه‌ها، مستلزم حفظ یک سطح حداقلی از «کنترل عملیاتی انسانی» است یا خیر؟ به‌عبارت‌دیگر، آیا «مراقبت» صرفاً به معنای به‌کارگیری بهترین مشاوره فنی در مرحله برنامه‌ریزی است، یا الزاماً متضمن نظارت یا امکان مداخله انسانی در حین اجرای عملیات توسط سامانه وجود دارد؟

پاسخ به این پرسش، نیازمند بررسی این موضوع است که آیا چارچوب مفهومی «کنترل انسانی معنادار»[1] - که در دکترین حقوقی برای پر کردن خلأهای نظارتی فناوری‌های خودکار پیشنهاد شده است - می‌تواند محتوای تکلیف مراقبت دائمی را در رابطه با سامانه‌های یادگیرنده روشن ساخته و معیاری برای «عملی بودن» مراقبت از آن‌ها ارائه دهد.

2-4. نقش کنترل انسانی مؤثر در تحقق مراقبت دائمی در استفاده از سامانه‌های خودکار

همان‌طور که پیش از این بیان شد، ابهام ذاتی در تفسیر تکلیف «مراقبت دائمی» در مواجهه با

1. MHC (Meaningful Human Control)؛ کنترلی انسانی معنادار بر سامانه‌های خودمختار، به‌طوری که انسان بتواند تصمیمات کلیدی سامانه را هدایت و اصلاح کند.

فناوری‌های نوین، این پرسش را برجسته می‌سازد که اعمال کنترل انسانی تا چه حد می‌تواند به تضمین اجرای مؤثر این تکلیف کمک کند؟ در پاسخ، دکترین حقوقی شاهد ظهور تحلیل‌ها و چارچوب‌های متفاوتی برای تبیین رابطه انسان، سامانه‌های خودکار و الزامات IHL بوده است.

2-4-1. دکترین «کنترل انسانی معنادار» (MHC): از مفهوم تا کاربرد

اصطلاح «کنترل انسانی معنادار» نخستین بار در گزارش سال ۲۰۱۳ سازمان غیردولتی[1] Article 36 درباره رویکرد بریتانیا نسبت به سامانه‌های تسلیحاتی خودکار مطرح شد (ARTICLE 36, Killer Robots: UK Government Policy on Fully Autonomous Weapons, 2013). این اصطلاح تاکنون در حد یک مفهوم دانشگاهی و نظری باقی مانده و در قوانین یا معاهدات بین‌المللی الزام‌آور جای نگرفته است. این دکترین،[2] متضمن چهار عنصر کلیدی برای اطمینان از مشروعیت یک سامانه تسلیحاتی خودکار است: (۱) پیش‌بینی پذیری، قابل‌اعتماد بودن و شفافیت؛ (۲) امکان تحصیل اطلاعات دقیق درباره نتیجه مورد انتظار و زمینه عملیاتی؛ (۳) قابلیت مداخله و اقدام انسانی به‌موقع؛ و (۴) امکان انتساب مسئولیت حقوقی به نتایج حاصله (Roff & Moyes, 2016).

در خصوص جایگاه این دکترین در نظام حقوقی، دو دیدگاه اصلی وجود دارد. دیدگاه حداکثری MHC را به‌عنوان یک اصل مستقل و مکمل در کنار اصول بنیادین IHL قرار می‌دهد. در مقابل، دیدگاه حداقلی آن را یک الزام حقوقی مستقل

---

1. NGO (Non-Governmental Organization): سازمان غیردولتی که به‌صورت مستقل از دولت‌ها فعالیت می‌کند و معمولاً در زمینه پژوهش، سیاست‌گذاری یا نظارت بر مسائل اجتماعی و حقوقی فعال است.

2. AWS (Autonomous Weapon Systems): سامانه‌های تسلیحاتی خودمختار که می‌توانند بدون دخالت مستقیم انسانی اهداف را شناسایی، تصمیم‌گیری و اقدام کنند.

ندانسته، بلکه صرفاً اصل راهنمایی برای طراحی و به‌کارگیری سامانه‌ها جهت تسهیل رعایت IHL می‌داند (Horowitz; Scharre; Center for a New American Security, 2015: 7). با توجه به بی‌میلی کنونی بسیاری از دولت‌ها به توافق بر سر قواعد جدید الزام‌آور حقوقی، رویکرد حداقلی واقع‌بینانه‌تر به نظر می‌رسد (Bhuta; Beck; Geiss, 2016: 375). با این حال، همین دیدگاه حداقلی نیز به نوبه خود به شقوق مختلفی تقسیم می‌شود که بر سر «میزان» و «لحظه» ضروری کنترل انسانی اختلاف‌نظر دارند.

۲-4-۲. دیدگاه الزام به دخالت پیشینی انسان: کنترل در سطح تاکتیکی

یکی از شاخه‌های اصلی در دیدگاه حداقلی، بر عنصر سوم دکترین MHC (مداخله انسانی به‌موقع) تمرکز ویژه دارد و استدلال می‌کند که بدون چنین مداخله‌ای، تکلیف مراقبت دائمی نقض می‌شود. برای نمونه، دیوید آکرسان استدلال می‌کند که تفویض اختیار تصمیم‌گیری نهایی حمله به یک سامانه خودمختار، شکافی بین نیروی نظامی مسئول رعایت تکلیف مراقبت و توانایی عملی او برای اعمال این تکلیف ایجاد کرده و بنابراین نقض تکلیف مراقبت دائمی محسوب می‌شود (Akerson, 2013: 87).

برای تعیین محل دقیق این مداخله، برخی نظریه‌پردازان به سطوح سه‌گانه جنگ متوسل می‌شوند: سطح راهبردی[1] (تعیین اهداف کلان)، سطح عملیاتی[2] (طرح‌ریزی و هماهنگی نبردها) و سطح تاکتیکی[3] (اجرای درگیری‌های فردی) (Curtis E. LeMay

---

1. Strategic: سطح راهبردی جنگ که اهداف کلان و سیاست‌های بلندمدت نظامی و امنیتی را تعیین می‌کند.
2. Operational: سطح عملیاتی که شامل برنامه‌ریزی و هدایت عملیات نظامی برای دستیابی به اهداف راهبردی است.
3. Tactical: سطح تاکتیکی که مربوط به مدیریت مستقیم نیروها و تجهیزات در میدان نبرد برای اجرای عملیات است.

(Center, 2019: 12). هدر روف و ریچارد مویس، با تمرکز بر این تمایز، استدلال می‌کنند که کنترل انسانی معنادار حداقل باید در سطح تاکتیکی اعمال شود. طبق استاندارد پیشنهادی آنان، یک سامانه تسلیحاتی خودکار نباید بتواند بدون نظارت و تأیید پیشینی یک انسان برای هر حمله مجزا، به‌طور خودکار از هدفی به هدف دیگر حرکت کند (Roff & Moyes, 2016: 4–5).

دلیل این امر از نظر روف و مویس، حفظ «آگاهی زمینه‌ای» است. آنان معتقدند اگر کنترل انسانی تنها به سطوح راهبردی یا عملیاتی (که از میدان نبرد فاصله فیزیکی و شناختی دارند) محدود شود، کیفیت نظارت و قضاوت‌های حقوقی و عملیاتی به‌تدریج کاهش می‌یابد. این فاصله شناختی می‌تواند به حدی برسد که توانایی پیش‌بینی دقیق پیامدهای عملیات غیرممکن یا به‌شدت محدود شود (Roff & Moyes, 2016: 5)؛ بنابراین، از این دیدگاه، تصویب انسانی در آستانه هر اقدام تهاجمی، شرط لازم برای تحقق عملی تکلیف مراقبت دائمی است.

3-۴-2. امکان ارتقای انطباق با حقوق بین‌الملل بشردوستانه از طریق سامانه‌های خودکار یادگیرنده بدون دخالت انسانی پیشینی

در مقابل دیدگاهی که کنترل انسانی پیشینی را برای تحقق تکلیف مراقبت دائمی ضروری می‌داند، استدلال متقابلی مطرح است که بر اساس آن، واگذاری تصمیمات حقوقی و عملیاتی به یک سامانه خودکار یادگیرنده نه‌تنها لزوماً کیفیت این تصمیمات را تضعیف نمی‌کند، بلکه در پاره‌ای شرایط می‌تواند بهبود بخش آن باشد. این دیدگاه بر دو محور اصلی استوار است: توانایی منحصربه‌فرد سامانه‌های یادگیرنده در پردازش

داده‌های کلان و تحلیل پیچیده و محدودیت‌های ذاتی انسان در نظارت مؤثر بر چنین سیستم‌هایی.

- نقش داده‌های کلان و قابلیت پردازش در بهبود تصمیم‌گیری

تصمیم‌گیری در عملیات نظامی معاصر به‌طور فزاینده‌ای داده محور شده و داده‌های کلان در آن نقشی حیاتی ایفا می‌کنند (Trapp, 2013: 159–160). این داده‌ها در تمام مراحل چرخه هدفگیری - از تعیین و اولویت‌بندی اهداف تا تحلیل توانمندی‌ها، تخصیص نیرو، اجرای مأموریت و ارزیابی نهایی - جریان دارند (Spiegeleire, Maas & Sweijs, 2017: 89). هر یک از این مراحل دارای حلقه بازخورد و تأخیرهای زمانی خاص خود است که نیاز به پردازش سریع و دقیق حجم عظیمی از اطلاعات را مضاعف می‌کند (Spiegeleire, Maas & Sweijs, 2017: 89). در چنین شرایطی، حجم داده می‌تواند توان تحلیلگران انسانی را تحت فشار قرار داده و به خطاهای شناختی منجر شود (Spiegeleire, Maas & Sweijs, 2017: 89). در مقابل، واگذاری وظایف تحلیل داده در این چرخه به سامانه‌های یادگیری ماشینی می‌تواند تصمیمات عملیاتی را بهبود بخشد. دلیل این امر، قابلیت به‌روزرسانی مستمر و بی‌وقفه این سامانه‌ها مطابق با واقعیت‌های متغیر میدانی است. ذات سامانه‌های یادگیرنده مبتنی بر بهره‌گیری از تجربه و بهبود خودکار عملکرد از طریق اصلاح ساختار، برنامه یا داده‌هاست (Boulanin & Verbruggen, 2017: 16)؛ بنابراین، می‌توان تصور کرد که یک سامانه تسلیحاتی خودکار یادگیرنده در آینده بتواند اطلاعات و تجربیات را در سطوح تاکتیکی و عملیاتی یکپارچه کرده و برنامه‌ها را بازطراحی کند. این قابلیت در تئوری می‌تواند سه سطح جنگ را تلفیق و تأخیرها و سوءتفاهم‌های ناشی از حلقه‌های بازخورد سنتی را حذف کند.

شواهد از حوزه‌های غیرنظامی نیز حاکی از پتانسیل فوق‌العاده هوش مصنوعی در تصمیم‌گیری‌های پیچیده است. برای نمونه، عملکرد سامانه‌هایی مانند AlphaGo Zero در بازی GO نشان داد که فناوری‌های یادگیری ماشینی پیشرفته می‌توانند در حیطه‌های پیچیده راهبردی، فراتر از توانایی‌های انسان عمل کنند (Kania, 2017). این نمونه، پتانسیل هوش مصنوعی برای ایفای نقشی اساسی در تصمیم‌گیری‌های آینده در درگیری‌های مسلحانه را به‌وضوح نشان می‌دهد (Kania, 2017).

با این توصیف، این پرسش مطرح می‌شود که سامانه‌های یادگیرنده آینده چگونه می‌توانند مستقیماً به افزایش انطباق با حقوق بین‌الملل بشردوستانه بینجامند؟ یک مثال عینی، استفاده از فناوری تشخیص چهره پیشرفته است. چنین سامانه‌ای می‌تواند با پردازش حجم وسیعی از داده‌های تصویری با سرعت و دقتی فراتر از انسان، رزمندگان را از غیرنظامیان تمییز دهد. خروجی این تحلیل سپس می‌تواند به‌عنوان مبنایی برای اتخاذ قضاوت‌های حقوقی دقیق‌تر - مانند ارزیابی خسارات جانبی احتمالی برای اعمال اصل تناسب - مورد استفاده قرار گیرد (Turner, 2018: 356).

با این حال، تحلیل فوق ناقص خواهد بود اگر به یک چالش بنیادین اشاره نکند: مسئله کیفیت و سوگیری داده‌ها. سامانه‌های یادگیرنده ذاتاً وابسته به داده‌هایی هستند که با آن‌ها آموزش می‌بینند و تغذیه می‌شوند. اگر داده‌های آموزشی ناقص، تاریخ گذشته یا حاوی سوگیری‌های سیستماتیک (مانند شناسایی نادرست گروه‌های خاص به‌عنوان اهداف نظامی) باشند، تصمیمات خروجی سامانه نیز معیوب و احتمالاً نقض‌کننده حقوق بشردوستانه خواهد بود. این وابستگی مطلق به داده، لزوم نظارت و اعتبارسنجی مداوم را -

حتی بر سامانه‌های بسیار پیشرفته - گوشزد می‌کند.

درعین‌حال، همین قابلیت‌های پردازشی بالا، معمای نظارت انسانی را پیش می‌کشد. روشن نیست که یک اپراتور انسانی - به‌ویژه در سناریوهای حساس به زمان - تا چه حد می‌تواند بر فرآیندهای سریع و پیچیده یک سامانه یادگیرنده نظارت معناداری اعمال کند. همان‌گونه که یکی از محققان خاطرنشان ساخته،[1] «باوجود برتری اطلاعاتی ماشین نسبت به اپراتور، دخالت انسانی معنادار امکان‌پذیر نیست... و به دلیل سرعت پردازش و تعدد متغیرهای عملیاتی، کنترل ماشین توسط انسان واقعیت نخواهد داشت» (Matthias, 2004:182–183). در چنین وضعیتی، اصرار صرف بر الزام تصویب پیشین انسانی - بدون در نظر گرفتن ماهیت این سامانه‌ها - ممکن است ناشی از درک نادرست از عملکرد آن‌ها باشد و لزوماً تضمینی برای افزایش رعایت حقوق بین‌الملل بشردوستانه ایجاد نکند.

- محدودیت ذاتی نظارت انسانی و ناکارآمدی الزام به تصویب پیشین

چالش نظارت انسانی هنگامی عمیق‌تر می‌شود که مسئله «قابل توضیح نبودن» یا «جعبه سیاه» بودن سامانه‌های یادگیرنده موردتوجه قرار گیرد. سامانه‌های تسلیحاتی خودکار یادگیرنده، حتی در صورت کار با داده‌های محدود، ممکن است از فرآیندهای پیچیده‌ای مانند یادگیری عمیق استفاده کنند که برای انسان به‌طور شفاف قابل‌درک نیست (Mittelstadt et al., 2016: 4, 6). نمونه بارز این مسئله، خودروی خودران و خودیاد گیر شرکت NVIDIA است که نه بر اساس دستورالعمل‌های صریح برنامه‌نویسی، بلکه با مشاهده و تحلیل رفتار

1. آندریاس متیاس / Andreas Matthias: پژوهشگر حوزه مسئولیت‌های قانونی سامانه‌های یادگیرنده و هوش مصنوعی.

رانندگی انسان آموزش دید (Knight, 2017). مشکل اصلی اینجاست که دقیقاً مشخص نیست این سامانه در لحظه تصمیم‌گیری چگونه عمل می‌کند. پیچیدگی سیستم به حدی است که حتی مهندسان طراح آن نیز ممکن است نتوانند دلیل یک اقدام خاص را تشخیص دهند و بنابراین راه روشنی برای توضیح منطق تصمیمات اتخاذشده وجود ندارد (Knight, 2017).

این ویژگی با سامانه‌های خودکار مبتنی بر قواعد ساده «اگر - آنگاه» که پیش‌تر مرسوم بودند در تقابل است. در سامانه‌های قدیمی‌تر، تشخیص ناهنجاری در فرآیند تصمیم‌گیری نسبتاً آسان بود و این ناهنجاری می‌توانست به‌عنوان هشداری برای دخالت و لغو دستور توسط ناظر انسانی عمل کند. در مقابل، پیچیدگی فناوری یادگیری ماشینی، شناسایی همین ناهنجاری‌ها را برای ناظر انسانی دشوار یا غیرممکن می‌سازد.

وزارت دفاع ایالات‌متحده این «هسته تاریک و مخفی هوش مصنوعی» را مانعی اساسی در به‌کارگیری نظامی سامانه‌های یادگیرنده دانسته است (Knight, 2017). این نهاد حتی برنامه‌ای با عنوان «هوش مصنوعی قابل توضیح» راه‌اندازی کرده که هدف آن توسعه روش‌هایی برای ارائه توضیح درباره تصمیمات سامانه‌های یادگیرنده است (Knight, 2017). با این حال، حتی این راه‌حل‌ها با محدودیت‌های جدی روبه‌رو هستند: نخست آنکه توضیحات ارائه‌شده معمولاً ساده‌سازی شده و ممکن است جزئیات حیاتی را حذف کنند (Mittelstadt et al., 2016: 4). دوم، تهیه و درک این توضیحات غالباً زمان‌بر است (Mittelstadt et al., 2016: 4-6) درحالی‌که در میدان نبرد، ثانیه‌ها می‌تواند سرنوشت‌ساز باشد.

درنتیجه، تصمیم‌گیری در سامانه‌های یادگیرنده بر پایه داده‌های کلان و فرآیندهای تحلیلی غیر شفاف استوار است. در چنین شرایطی، ادعای نظارت یا کنترل لحظه‌ای و معنادار انسان با تردیدهای جدی مواجه می‌شود. با این حال، پرسش

نهایی این است: آیا همین سامانه‌های پیچیده و غیرقابل‌فهم - و شاید دقیقاً به دلیل قابلیت‌های پردازشی فوق‌العاده‌شان - ممکن است در عمل بتوانند تصمیمات هدفگیری منطبق‌تری با حقوق بین‌الملل بشردوستانه نسبت به سامانه‌های تحت کنترل مستقیم و سنتی انسان اتخاذ کنند؟ اگر پاسخ به این پرسش در مواردی مثبت باشد، نمی‌توان صرفاً به دلیل دشواری نظارت معنادار انسانی، استفاده از آن‌ها را مطلقاً ممنوع کرد، مگر آنکه دلیل قطعی و اقامه‌شده‌ای بر ضرورت غیرقابل تفویض بودن چنین نظارتی در تمام حالات وجود داشته باشد. این بحث، محور اصلی مناقشه در بازتعریف مفهوم «کنترل انسانی» در عصر هوش مصنوعی است.

- نظریه «احتیاط پویا» مارگولیس و کنترل انسانی معنادار

در تقابل با دیدگاه راف و مویز که بر ناتوانی ذاتی کنترل انسانی در سطح تاکتیکی در مواجهه با پیچیدگی و سرعت سامانه‌های یادگیرنده تأکید می‌کنند، نسخه‌های دیگری از مفهوم کنترل انسانی معنادار که حداقلی از تصویب انسانی پیش از استفاده از نیرو را ضروری می‌دانند، در ادبیات علمی و در میان برخی از اعضای جامعه بین‌المللی طرفدارانی دارد (Goose, 2014). با این حال، برای انطباق با ماهیت پویا و غیرقابل‌پیش‌بینی سامانه‌های یادگیرنده، نیاز به بازتعریف این مفهوم احساس می‌شود. یکی از جایگزین‌های مفهومی پیشنهادی که هماهنگی بیشتری با این پیچیدگی دارد، استاندارد «احتیاط پویا»[1] ارائه‌شده توسط پیتر مارگولیس[1]

1. dynamic diligence (احتیاط پویا): استانداردی که توسط پیتر مارگولیس پیشنهاد شده و به توزیع منعطف کنترل بین انسان و ماشین در عملیات نظامی اشاره دارد، به‌طوری که دخالت انسانی تنها در حد ضروری و عملی برای کاهش خطرات غیرنظامیان اعمال

است (Margulies, 2017: 415–416). هسته اصلی این نظریه، توزیع انعطاف‌پذیر اختیار بین انسان و ماشین، هم در مرحله برنامه‌ریزی و هم در مرحله اجرا، بر اساس شرایط متغیر عملیاتی است (Margulies, 2017:, p. 433). این مدل دو رکن اصلی دارد:

۱. قابلیت سامانه تسلیحاتی برای درخواست بررسی و تأیید انسانی پیش از اقدام در محیط‌های پرخطر (مانند مناطق شهری با تراکم غیرنظامی)؛

۲. امکان لغو دستورها یا پروتکل‌های یادگیری ماشین توسط اپراتور انسانی در هر مرحله (Margulies, 2017: 433–434).

مارگولیس استدلال می‌کند که در برخی شرایط، دخالت انسانی ممکن است ضرورتی نداشته باشد (Margulies, 2017: 434). در وضعیت‌هایی که سرعت واکنش حیاتی است و سامانه بسیار سریع‌تر از انسان عمل می‌کند، الزام به تصویب قبلی انسانی برای هر اقدام تهاجمی نه‌تنها ممکن است «عملی» نباشد، بلکه می‌تواند دستیابی به هدف نظامی مشروع را با اختلال مواجه سازد (Margulies, 2017:434). وی همچنین با گزاره کلی که «نظارت انسانی همیشه به رعایت بهتر تکلیف مراقبت دائمی منجر می‌شود» مخالف است و معتقد است باید به تجربه عملی توجه کرد که آیا دخالت انسان لزوماً از تلفات غیرنظامیان پیشگیری می‌کند یا خود می‌تواند به عاملی برای افزایش خطر تبدیل شود (Margulies, 2017: 434). مارگولیس

شود و لزوماً همیشه دخالت انسان بهترین نتیجه را تضمین نمی‌کند.

1. Peter Margulies: استاد حقوق در دانشگاه، پژوهشگر حوزه حقوق بین‌الملل بشردوستانه و فناوری‌های نظامی نوین، به‌ویژه سامانه‌های تسلیحاتی خودمختار و رابطه بین کنترل انسانی و مسئولیت حقوقی. او استاندارد «احتیاط پویا» را برای توزیع منعطف کنترل بین انسان و ماشین در عملیات نظامی پیشنهاد کرده و بر این نکته تأکید دارد که دخالت انسانی لزوماً همیشه بهترین نتیجه را برای حفاظت از غیرنظامیان به همراه ندارد.

خاطرنشان می‌سازد که حتی در مواردی که تصویب انسانی از نظر فنی امکان‌پذیر است، این تصویب لزوماً به بهبود رعایت حقوق بین‌الملل بشردوستانه منجر نمی‌شود؛ زیرا ممکن است در هدف نهایی تکلیف مراقبت دائمی - که «حفاظت از جمعیت غیرنظامی و اهداف غیرنظامی» است (پروتکل الحاقی اول، ۱۹۷۷، ماده ۵۷) - اختلال ایجاد کند. وی این‌گونه جمع‌بندی می‌کند: «اگرچه سامانه‌های تسلیحاتی خودکار باید قابلیت دخالت انسانی را داشته باشند، اما اگر بتوانند وظیفه‌ای را به‌صورت مستقل با همان کیفیت یا حتی بهتر از حالتی که تحت همراهی انسان انجام می‌شود انجام دهند، حقوق بین‌الملل بشردوستانه الزامی برای دخالت انسانی تحمیل نمی‌کند» (Margulies, 2017: 434). دلیل این امر آن است که انکارناپذیر است ماشین به جهت مصونیت از خطاها و اعوجاج‌های قضاوتی ناشی از هیجانات و محدودیت‌های شناختی انسان، می‌تواند عملکردی دقیق‌تر و قابل پیش‌بینی‌تر ارائه دهد.

با این حال، حتی این رویکرد پیشرفته‌تر نیز برای سامانه‌های یادگیرنده آینده ممکن است کافی نباشد. در مورد سلاح‌های خودکار ساده‌تر با منطق شرطی «اگر - آنگاه»، به دلیل شفافیت نسبی فرآیند و پیش‌بینی پذیری رفتار، امکان نظارت و دخالت انسانی منطقی به نظر می‌رسد (Boulanin & Verbruggen, 2017: 9)؛ اما در مورد سامانه‌های یادگیرنده پیچیده‌تر، دستیابی به نتایج بهینه مطابق با حقوق بین‌الملل بشردوستانه ممکن است مستلزم فراتر رفتن از چارچوب مارگولیس و در نظر گرفتن سناریوهایی باشد که در آن‌ها امکان لغو انسانی وجود ندارد. برای درک این امکان، می‌توان به فناوری یادگیری تقویتی AlphaGo Zero اشاره کرد که نسخه‌ای قوی‌تر از AlphaGo بود. این سامانه، آزاد از «تجربیات قبلی، قواعد مرسوم و خرد متعارف که

اغلب تصمیم‌گیرندگان انسانی بر آن تکیه می‌کنند» (Scherer, 2016: 365)، صرفاً با تمرین در بازی با خود، موفق به کشف استراتژی‌های کاملاً غیرمتعارف و حرکات خلاقانه‌ای شد که توسط انسان‌ها پیش‌بینی نشده بود (Silver & Hassabis, 2017). به همین ترتیب، یک سامانه تسلیحاتی یادگیرنده آینده که از الگوریتم‌های مشابه استفاده کند، ممکن است به تصمیمات هدف‌گیری‌ای دست یابد «که انسان‌ها هرگز به آن فکر نکرده یا آن را رد کرده و گزینه‌های شهودی جذابتری را انتخاب کرده باشند» (Scherer, 2016: 365).

تفسیر تکلیف مراقبت دائمی همواره باید با توجه به هدف غایی ماده ۵۷ پروتکل الحاقی اول - یعنی «حفاظت از جمعیت غیرنظامیان و اهداف غیرنظامی» - صورت پذیرد (پروتکل الحاقی اول، ۱۹۷۷، ماده ۵۷)؛ بنابراین، الگوی ساده انگارانه «هرچه کنترل انسانی بیشتر، بهتر» که در برخی روایت‌های کنترل انسانی معنادار دیده می‌شود، باید در پرتو قابلیت‌های نوظهور فناوری‌های یادگیری ماشینی مورد بازبینی انتقادی قرار گیرد. پرسش محوری باید این باشد: آیا نظارت یا دخالت انسانی در یک موقعیت مشخص، واقعاً منجر به کاهش آسیب به غیرنظامیان می‌شود؟

- نظریه «پیش‌بینی معقول» شولر[1]

همان‌طور که تحلیل شد، فرمول‌بندی‌های راف و مویز و نیز مارگولیس از استاندارد کنترل

---

1. Schuller's Reasonable Predictability Theory: نظریه‌ای که توسط آلن شولر مطرح شده و بر این اساس، سامانه‌های یادگیرنده خودمختار نباید از سطحی بالاتر از «پیش‌بینی معقول» برای رعایت قوانین بشردوستانه بین‌المللی عبور کنند. منظور از «پیش‌بینی معقول» این است که عملکرد سامانه باید تا حد امکان قابل پیش‌بینی باشد و احتمال رعایت قوانین IHL توسط آن قابل‌سنجش باشد، بدون اینکه انتظار برآورده شدن کامل رعایت مطلق قوانین وجود داشته باشد. این نظریه به تعیین حداقل سطح مداخله انسانی و میزان قابل‌قبول مسئولیت قانونی کمک می‌کند.

انسانی معنادار ناکافی به نظر می‌رسند، زیرا یا پتانسیل سامانه‌های یادگیرنده را نادیده می‌گیرند یا آن را محدود می‌کنند. درواقع، ممکن است رعایت بهتر حقوق بین‌الملل بشردوستانه یا عملکرد نظامی بهینه، مستلزم عملیاتی باشد که در آن دخالت یا کنترل مستقیم انسانی در حلقه اقدام وجود نداشته باشد.

نسخه ارائه‌شده توسط شولر از دکترین کنترل انسانی معنادار، فاصله معناداری با مفهوم سنتی کنترل به‌مثابه نظارت انسانی لحظه‌ای می‌گیرد (Schuller, 2017). در عوض، آزمون وی مبتنی بر این معیار است: آیا یک سامانه تسلیحاتی خودکار می‌تواند با سطح معقولی از اطمینان پیش‌بینی شود که با حقوق بین‌الملل بشردوستانه سازگار عمل خواهد کرد؟ (آزمون «پیش‌بینی معقول») (Schuller, 2017: 408–409). بر این اساس، اگر اپراتور انسانی استفاده‌کننده از یک سامانه یادگیرنده تأیید کند که سامانه مذکور آزمون پیش‌بینی معقول را پاس کرده است، دیگر نیازی به هیچ‌گونه تعامل یا تصویب اضافی انسانی پیش از اقدام مرگبار وجود ندارد (Schuller, 2017: 420–421).

می‌توان گفت رویکرد شولر در انتقال کانون توجه از «کنترل مداخله‌گرایانه انسانی» به «کنترل مبتنی بر پیش‌بینی پذیری»، در میان نسخه‌های بررسی‌شده، بیشترین سازگاری را با ماهیت سامانه‌های یادگیرنده آینده دارد. برخلاف مارگولیس که هنوز برای دخالت انسانی در شرایط خاص مجالی قائل است، شولر در شرایطی که پیش‌بینی معقولی از انطباق سامانه با حقوق بین‌الملل بشردوستانه وجود داشته باشد، با تفویض کامل اختیار به سامانه و جداسازی آن از کنترل مستقیم انسانی در حین عمل مخالفتی ندارد (Schuller, 2017: 420–423).

با وجود این مزیت، استاندارد پیش‌بینی معقول شولر نیز خالی از اشکال نیست. این استاندارد عمدتاً بر نتیجه (انطباق با حقوق بین‌الملل بشردوستانه) متمرکز است و راهنمای عملی روشنی برای چگونگی توسعه و اعتبارسنجی سامانه‌های تسلیحاتی خودکار قانونی ارائه نمی‌دهد (Schuller, 2017: 415–425). شولر تنها به اصولی کلی بسنده می‌کند:

۱. سامانه‌های تسلیحاتی خودکار می‌توانند صرفاً از طریق برنامه‌نویسی به شیوه‌ای قانون‌مدار کنترل شوند.
۲. حقوق بین‌الملل بشردوستانه، تصویب انسانی فوری پیش از هر اقدام مرگبار را الزامی نمی‌داند.
۳. الزام پیش‌بینی معقول لازم نیست در مورد تمامی جنبه‌های رفتار سامانه اعمال شود، بلکه تنها در موارد مرتبط با رعایت حقوق بین‌الملل بشردوستانه کفایت می‌کند.
۴. ظرفیت تخریبی سامانه را می‌توان از طریق محدودیت‌های فیزیکی و برنامه‌نویسی مهار کرد (Schuller, 2017: 417–425).

درنهایت، شولر تأکید می‌کند که تصمیم هدف‌گیری مرگبار نباید به‌گونه‌ای به رایانه واگذار شود که اطمینان انسانی به رعایت استاندارد پیش‌بینی معقول توسط ماشین را خدشه‌دار کند (Schuller, 2017: 415–425). با وجود برخی ناسازگاری‌های این اصول با ویژگی‌های کلیدی سامانه‌های یادگیرنده (مانند غیرقابل‌پیش‌بینی بودن ذاتی)، دکترین شولر نقطه آغاز مناسبی برای تدوین چارچوبی در مورد چگونگی اعمال تکلیف مراقبت دائمی بر این فناوری‌ها فراهم می‌آورد. بخش بعدی مقاله، با اتکا بر این مبانی، اصول تکمیلی را برای توسعه سامانه‌های تسلیحاتی خودکار یادگیرنده در چارچوب قانونی پیشنهاد خواهد داد.

- محدودیت‌ها و چالش‌های کنترل انسانی در سامانه‌های یادگیرنده خودکار

شولر استاندارد «پیش‌بینی معقول» خود را با استناد به ذاتِ همراه با درجاتی از عدم قطعیت ماشین‌های یادگیرنده توجیه می‌کند (Schuller, 2017: 409–413). با این حال، وی به وضوح توضیح نمی‌دهد که این استاندارد در شرایطی که کنترل بین سامانه تسلیحاتی خودکار و اپراتور انسانی تقسیم شده است، چگونه باید اعمال شود. او صرفاً بیان می‌کند؛ «اگر اپراتور انسانی نتواند مطابقت ماشین با حقوق بین‌الملل بشردوستانه را پیش‌بینی کند، خودمختاری آن ممکن است غیرقانونی باشد» (Schuller, 2017: 409). این در حالی است که در سناریوی تقسیم کنترل، انطباق نهایی با حقوق بین‌الملل بشردوستانه به عملکرد هر دو عامل، یعنی ماشین و اپراتور انسانی، وابسته است.

در عمل، شرایط مختلفی قابل‌تصور است که در آن‌ها کنترل مستقیم انسانی بر عملکرد یک سامانه یادگیرنده اعمال می‌شود.[1] برای درک این طیف، می‌توان به چارچوب سطوح خودکارسازی در صنعت خودروهای خودران[2] مراجعه کرد. انجمن استاندارد گذاری حرفه‌ای مهندسان خودرو،[3] سطوح

---

1. «Human-on-the-loop» به معنای وضعیتی است که در آن یک سامانه خودمختار یا نیمه خودمختار قادر به انجام عملیات است، اما انسان نظارت کلی و کنترل راهبردی بر عملکرد آن دارد و می‌تواند در صورت لزوم مداخله کند یا عملیات را متوقف سازد. برخلاف «human-in-the-loop» که حضور مستقیم انسان برای تصمیم‌گیری الزامی است، در «human-on-the-loop» انسان بیشتر نقش نظارت‌کننده و اصلاح‌کننده را ایفا می‌کند تا تصمیم‌گیرنده لحظه‌ای. این مفهوم در بحث سامانه‌های تسلیحاتی خودمختار و حقوق بین‌الملل بشردوستانه اهمیت دارد؛ زیرا سطح دخالت انسانی می‌تواند تأثیر مستقیمی بر مسئولیت حقوقی و رعایت تکالیف مراقبت دائمی داشته باشد.

2. SAE International: یک سازمان بین‌المللی استانداردسازی در زمینه مهندسی خودرو و هوافضا است.

3. پیشتر با نام انجمن مهندسین خودرو و هوافضا شناخته می‌شد.

مختلفی را تعریف کرده است: در سطح خودکارسازی جزئی (سطح ۲)، اپراتور انسانی به‌طور دائم محیط رانندگی را نظارت می‌کند و در صورت لزوم (بر اساس قضاوت خود) فرمان می‌دهد، شتاب می‌گیرد یا ترمز می‌کند (Smith, 2016: 98). در مقابل، در سطح خودکارسازی شرطی (سطح ۳)، نظارت بر محیط به عهده وسیله نقلیه است و کنترل انسانی تنها در مواقع ضروری و در پاسخ به درخواست سیستم صورت می‌پذیرد (Smith, 2016: 98).

در سطوح پایین‌تر خودکارسازی مانند سطح ۲، پدیده روان‌شناختی نگران‌کننده‌ای به نام «سوگیری خودکارسازی»[1] می‌تواند رخ دهد. این پدیده به تمایل انسان به اعتماد افراطی به سیستم خودکار، علیرغم وجود شواهدی مبنی بر نادرست یا غیرقابل‌اعتماد بودن آن در یک موقعیت خاص، اشاره دارد و نگرانی از واگذاری بیش‌ازحد مسئولیت به سیستم توسط کاربران را برمی‌انگیزد (Grut, 2013:14–15; Cummings, 2006).

این پدیده منحصر به خودروهای خودران یا سامانه‌های تسلیحاتی خودکار یادگیرنده نیست و در حوزه‌های دیگر مانند هدایت هواپیما نیز مشاهده شده است (Mosier et al., 2001). درواقع، هر زمان که ماشین‌ها در فرآیند تصمیم‌گیری انسانی نقش ایفا می‌کنند، امکان بروز این سوگیری وجود دارد (Grut, 2013: 14–15). سیستم‌های یادگیری ماشینی ممکن است این مشکل را از دو طریق تشدید کنند (Gretton, 2017). نخست، آگاهی از پیچیدگی الگوریتم‌های یادگیری، توأم با غیرقابل فهم بودن فرآیند تصمیم‌گیری آن‌ها (همان «هسته تاریک هوش مصنوعی»)، می‌تواند منجر به افزایش

---

1. Automation bias: تمایل انسان‌ها به اعتماد بیش‌ازحد به پیشنهادها یا تصمیم‌های سیستم‌های خودکار، حتی زمانی که اشتباه می‌کنند. این پدیده می‌تواند منجر به خطاهای انسانی شود؛ زیرا اپراتورها فرض می‌کنند سیستم‌های خودکار همیشه درست عمل می‌کنند.

تمایل انسان به اعتماد کورکورانه به ماشین شود (Grut, 2013: 19) تا آنجا که حتی با آشکار شدن نشانه‌های خرابی سیستم نیز اپراتور مداخله نکند (Neslage, 2015: 173–174). دوم، گاهی مداخله انسانی خود می‌تواند به نتایجی بدتر از نظر انطباق با حقوق بین‌الملل بشردوستانه بینجامد که این امر به نوبه خود می‌تواند به‌صورت معکوس، موجب اجتناب اپراتور از مداخله لازم در موقعیت‌های دیگر شود.

تمرکز شولر بر عدم قطعیت ناشی از تعامل سامانه یادگیرنده با محیط پیچیده میدان نبرد، استاندارد پیش‌بینی معقول او را با ضعفی مهم مواجه ساخته است (Schuller, 2017: 409–413)؛ زیرا در این چارچوب، به پیش‌بینی ناپذیری ناشی از تعامل دوسویه و پویای سامانه یادگیرنده با اپراتور انسانی توجه کافی نشده است. درحالی‌که برای بازتاب کامل ریسک‌های واقعی «سوگیری خودکار سازی» یا «سوگیری معکوس خودکار سازی»، یک رویکرد سه‌بعدی ضروری است. منبع پیش‌بینی‌ناپذیری تنها تعامل بین سامانه یادگیرنده و محیط عملیاتی نیست؛ بعد سوم و حیاتی، تعامل بین سامانه یادگیرنده و اپراتور انسانی است که خود می‌تواند منشأ رفتارهای غیرقابل انتظار باشد.

بنابراین، در فازهای طراحی، آزمون و ارزیابی، یک سامانه تسلیحاتی خودکار یادگیرنده باید نه‌تنها در معرض سناریوهای متنوع میدان نبرد (مانند شرایط مختلف جغرافیایی، آب‌وهوایی، تراکم و موقعیت غیرنظامیان و رزمندگان) قرار گیرد، بلکه ضروری است با اپراتورهای انسانی متفاوت نیز به تمرین و تعامل بپردازد تا الگوهای رفتاری و نتایج این تعامل انسان-ماشین ارزیابی شود. این رویکرد به اطمینان از لحاظ شدن کامل ملاحظات روان‌شناختی و پیامدهای غیرقابل‌پیش‌بینی

احتمالی ناشی از تعامل پیچیده انسان و سامانه هوشمند کمک شایانی می‌کند.

- فراتر از پیش‌بینی معقول به‌عنوان حداقل معیار قانونی و ضرورت ارتقای رعایت حقوق بین‌الملل بشردوستانه

در آزمون پیشنهادی شولر، نیازی به حصول اطمینان قطعی و کامل از رعایت حقوق بین‌الملل بشردوستانه - که عموماً هدفی دست‌نیافتنی تلقی می‌شود - نیست (Schuller, 2017: 408; نگاه کنید به Prosecutor v. Delalić, Case No. IT-96-21-T, Judgment: para. 395, 1998)؛ اما وی استدلال می‌کند که هر استانداردی که سطح پایین‌تری از «پیش‌بینی معقول» را بپذیرد، ممکن است موجب شود اپراتورهای انسانی، کاستی‌های سامانه‌های یادگیرنده را بهانه‌ای برای تخطی از تعهدات خود قلمداد کنند (Schuller, 2017: 408). شولر همچنین توضیح می‌دهد که استاندارد «معقول بودن» به‌عنوان یک معیار حقوقی شناخته‌شده و جاافتاده، این مزیت را دارد که سطوح بالاتر اطمینان را به دلیل پیچیدگی ذاتی برنامه‌نویسی رایانه‌ای و تشدید آن توسط «مه» میدان نبرد، غیرقابل دسترس می‌داند (Schuller, 2017: 408).

با این حال، این تحلیل تنها بخشی از واقعیت را بازتاب می‌دهد. درحالی‌که «قابلیت پیش‌بینی معقول» می‌تواند به‌عنوان حداقل معیار برای تعیین قانونی بودن خودمختاری یک سامانه تسلیحاتی خودکار یادگیرنده در نظر گرفته شود، متناظر با آن، باید تکلیفی ممکن و عملی برای بهینه‌سازی و ارتقای هرچه بیشتر سطح رعایت حقوق بین‌الملل بشردوستانه توسط چنین سامانه‌هایی نیز وجود داشته باشد. در این زمینه، تفسیر غایی و پیشرو از تکلیف «مراقبت دائمی» نه‌تنها ایجاب می‌کند که سامانه تسلیحاتی یادگیرنده به نتایجی دست یابد که رعایت حقوق بین‌الملل بشردوستانه در آن‌ها

به‌طور معقول قابل پیش‌بینی است؛ بلکه فراتر از این حداقل استاندارد، لازم است به‌گونه‌ای طراحی، آزمون و به کار گرفته شود که احتمال رعایت حقوق بین‌الملل بشردوستانه را به حداکثر ممکن برساند؛ بنابراین، یک استاندارد اصلاح‌شده و جامع باید سطحی از قابلیت پیش‌بینی و عملکرد را تضمین کند که هم تا حد امکان بهینه باشد و هم حداقل با معیار پایه «پیش‌بینی معقول» سازگاری داشته باشد. این رویکرد، گذار از یک چارچوب صرفاً دفاعی- حقوقی به سمت یک چارچوب ایجابی - کارکردی را نشان می‌دهد.

4-۴-2. ضرورت رعایت قابلیت پیش‌بینی بهینه در انطباق با حقوق بین‌الملل بشردوستانه

شولر استدلال می‌کند که دولت‌ها استانداردی فراتر از «معقول بودن» را نمی‌پذیرند؛ استانداردی که به‌طور گسترده در حقوق بین‌الملل بشردوستانه پذیرفته و درک شده است و «برای مدتی طولانی منافع متضاد حقوق بین‌الملل بشردوستانه را متوازن کرده است» (Schuller, 2017: 409). با این حال، هدف از طرح «قابلیت پیش‌بینی بهینه» افزایش یا تغییر استاندارد حقوقی موجود نیست؛ بلکه غایت آن، اطمینان از این امر است که تکلیف مراقبت دائمی به‌درستی و در کنار سایر تکالیف احتیاطی که محتوای آن را عملیاتی می‌سازند - مانند اصل حداقل خسارت پیش‌بینی‌شده جانبی مطابق ماده ۵۷ (۲) (الف) (ii) و تکلیف انتخاب دقیق اهداف مطابق ماده ۵۷ (۳) پروتکل الحاقی اول به کنوانسیون‌های ژنو مورخ ۱۲ اوت ۱۹۴۹ - اعمال شود (Additional Protocol I, 1977, arts. 57(2)(a)(ii), 57(3); Dill, 2010: 9)[1].

---

1. مواد ۵۷ پروتکل الحاقی اول، به اصول تناسب و احتیاط در عملیات نظامی اشاره دارد و چارچوبی برای ارزیابی میزان مجاز حمله به اهداف نظامی ارائه می‌کند.
ماده ۵۷ – اقدامات احتیاطی در حملات (Article 57 – Precautions in attack)

هر دوی این قواعد، بازتاب‌دهنده اصل ضرورت نظامی هستند که تنها اقداماتی را مجاز می‌داند که برای دستیابی به یک هدف نظامی مشروع ضروری باشند (af Jochnick & Normand, 1994: 49). اصل ضرورت نظامی، پایه حقوقی گسترده و مستحکمی برای آزمون «قابلیت پیش‌بینی بهینه» فراهم می‌آورد؛ زیرا هرگونه آسیب و خسارتی که فراتر از نیاز برای تضعیف توان دشمن باشد را منع می‌کند (Hague Convention IV, 1907: art. 23(g)).[1] علاوه بر این، همان‌گونه که پیشتر اشاره شد، اصل ضرورت نظامی به‌عنوان یکی از اصول بنیادین حقوق بین‌الملل بشردوستانه، زیربنای تفسیر کل این رشته حقوقی را تشکیل می‌دهد (Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996: para. 493).[2] بنابراین روشن است که اعمال صحیح این اصل در مورد یک سامانه تسلیحاتی خودکار یادگیرنده، مستلزم برنامه‌ریزی و طراحی آن سامانه به‌گونه‌ای است که حقوق بین‌الملل بشردوستانه را به‌طور کامل و در بالاترین سطح عملی ممکن از پیش‌بینی پذیری رعایت کند.

بند ۲ (الف) (ii): «در تمام عملیات حمله، باید اقدامات لازم برای کاهش احتمال خسارت جانبی بیش‌ازحد به غیرنظامیان و اموال غیرنظامی انجام شود. این اقدامات شامل انتخاب وسایل و روش‌های حمله‌ای است که بیشترین انطباق را با اهداف نظامی داشته و کمترین آسیب جانبی پیش‌بینی‌شده را ایجاد کند.»
بند ۳: «طرف‌های درگیر باید دقت کافی را در انتخاب اهداف نظامی داشته باشند و از هر هدفی که با اصول انسانی و ضرورت نظامی ناسازگار باشد، اجتناب کنند. در تعیین اهداف، باید تمام اطلاعات موجود و امکانات عملی را در نظر گرفت تا خسارت به غیرنظامیان و اموال غیرنظامی به حداقل برسد.»
به بیان ساده: ماده ۵۷ (۲) (الف) (ii) بر اصل کمینه کردن خسارت جانبی پیش‌بینی‌شده تأکید دارد و ماده ۵۷ (۳) بر انتخاب دقیق و هوشمندانه اهداف نظامی با رعایت اصول IHL تمرکز می‌کند.

1. اموال دشمن نمی‌توانند توقیف یا نابود شوند مگر آنکه «ضرورت‌های حتمی جنگ» ایجاب کند.
2. که بیان می‌دارد اصول بنیادین حقوق بین‌الملل بشردوستانه «هم به رشد و توسعه حقوق کمک می‌کنند و هم به‌عنوان تکیه‌گاهی برای هنجارها و آداب جامعه عمل می‌کنند».

5-۴-2. امکان‌پذیری عملی قابلیت پیش‌بینی بهینه در سامانه‌های خودکار یادگیرنده

یکی از استدلال‌های احتمالی علیه استاندارد قابلیت پیش‌بینی بهینه این است که تکالیف احتیاطی تنها آنچه را که «عملی» است ایجاب می‌کنند و مطالبه چیزی فراتر از «قابلیت پیش‌بینی معقول» خارج از حدود این تکلیف خواهد بود. مطابق با آنچه پیش‌تر در مورد ماهیت و هدف وصف «عملی بودن» بیان شد، در نسخه بازنگری شده راهنمای تالین[1] درباره قانون بین‌الملل حاکم بر عملیات سایبری، «عملی بودن» به معنای آنچه «با توجه به تمامی شرایط حاکم در آن زمان، ازجمله ملاحظات انسانی و نظامی، قابل انجام یا عملاً ممکن است» تعریف شده است (Tallinn Manual 2.0, 2017: 479). ازاین‌رو، می‌توان آن را به‌عنوان «اختیار عملیاتی و قلمرویی برای تفسیر» در نظر گرفت (Bhuta, Beck, & Geiss, 2016: 373) که اذعان دارد برای یک فرمانده انسانی در میدان نبرد فعال، دنبال کردن دقیق و لحظه‌به‌لحظه بهترین نتیجه ممکن از منظر حقوق بین‌الملل بشردوستانه، الزاماً عملی نیست (Bhuta, Beck, & Geiss, 2016: 376). مطالبه هر استاندارد بالاتر، می‌تواند «به‌گونه‌ای نامناسب، خطرات را به سمت سربازان منتقل کند که در سناریوهای پرخطر نباید با وظیفه اضافی ارزیابی امکان‌پذیری روش‌های کم‌خطرتر مواجه شوند» (Bhuta, Beck, & Geiss, 2016: 376).

با این حال، در مورد سامانه‌های تسلیحاتی خودکار، معیار «عملی بودن» منحصراً به

---

1. Deliberative system: سیستمی است که در آن تصمیم‌گیری‌های پیچیده، به‌ویژه در سامانه‌های یادگیرنده یا سامانه‌های تسلیحاتی خودمختار به‌صورت مرحله‌ای و منطقی انجام می‌شود و امکان ارزیابی گزینه‌ها و پیامدها پیش از اقدام فراهم می‌شود. این نوع سیستم‌ها شبیه یک فرآیند «تفکری» هستند که میان تحلیل، ارزیابی و اقدام فاصله ایجاد می‌کنند تا تصمیم نهایی بر اساس اطلاعات و ملاحظات متعدد اتخاذ شود.

محدودیت‌های قابلیت‌های انسانی مقید نیست. توانایی ذاتی ماشین می‌تواند و باید موجب شود سطح رعایت حقوق بین‌الملل بشردوستانه در عمل ارتقا یابد؛ زیرا «یک سلاح خودمختار می‌تواند وسیله‌ای برای عملی کردن برخی اقدامات احتیاطی باشد که برای یک سرباز ممکن نیست» (ICRC, Report on International Humanitarian Law and the Challenges of Contemporary Armed Conflicts, 2014: 42). برای مثال، سامانه‌های یادگیری ماشینی می‌توانند بر اساس «نظریه مطلوبیت»[1] برنامه‌ریزی شوند تا بهترین نتایج را در پیگیری یک هدف مشخص تولید کنند (ICRC, Report on International Humanitarian Law and the Challenges of Contemporary Armed Conflicts, 2014: 42). این نظریه با دستور دادن به ماشین برای عمل بر اساس احتمال وقوع برخی نتایج و تابع مطلوبیت آن نتایج در رسیدن به هدف موردنظر کار می‌کند. به‌عبارت‌دیگر، تصمیم منطقی چنین سامانه‌ای «به اهمیت نسبی اهداف مختلف و احتمال و درجه تحقق آن‌ها بستگی دارد» (Schuller, 2017: 411). سپس ماشین مسیر عملی‌ای را دنبال می‌کند که بهترین نتیجه را بر اساس این محاسبات فراهم می‌آورد (Schuller, 2017: 411).

فرض کنیم که برنامه‌ریزی یک سامانه تسلیحاتی خودکار یادگیرنده برای رعایت حقوق بین‌الملل بشردوستانه در سطح «قابل پیش‌بینی معقول» امکان‌پذیر باشد. بر اساس ملاحظات فوق، برنامه‌ریزی چنین سامانه‌ای برای دستیابی به هدف نظامی خاص یا انتخاب مسیر عملیاتی که بیشترین احتمال رعایت حقوق بین‌الملل بشردوستانه را دارد کاملاً «عملی» است. افزودن الگوریتم‌های لازم در مرحله طراحی و برنامه‌ریزی برای اجرای نظریه مطلوبیت، امری است که با صرف هزینه و تلاش نسبتاً معقولی قابل تحقق است. درواقع، بسیاری از سامانه‌های یادگیری ماشینی

1. Utility theory

کنونی، مانند آن‌هایی که از «یادگیری تقویتی»[1] استفاده می‌کنند، هم‌اکنون برای بهینه‌سازی نتایج بر اساس همین منطق برنامه‌ریزی شده‌اند (Stensmo & Sejnowski, 1996: 1061–1067). این امر کاملاً در چارچوب آنچه «قابل انجام یا عملاً ممکن» تلقی می‌شود قرار می‌گیرد (Henckaerts & Doswald-Beck, 2005: 54).

توانایی سامانه‌های یادگیرنده در ایجاد تعادل میان قواعد و ارزش‌های مختلف انسانی و اتخاذ تصمیم‌های پیچیده و درعین‌حال قابل دفاع - که منطبق بر ارزیابی و سنجش قوانین حقوقی، اخلاقی و اخلاقی- اجتماعی باشد - پیش‌تر در حوزه‌هایی مانند زیست اخلاق نشان داده شده است (Margulies, 2017: 420). مارگولیس تأکید می‌کند که این اصول منطقی می‌توانند به حوزه حقوق بین‌الملل بشردوستانه نیز تعمیم یابند (Margulies, 2017: 420).

برخلاف سامانه‌های خودکار ساده‌تر، توانایی روزافزون سامانه‌های یادگیرنده در انجام تصمیم‌گیری‌های دقیق و راهبردی، امکان انتقال تدریجی اختیار تصمیم‌گیری در چرخه هدف‌گیری به این سامانه‌ها را در آینده فراهم می‌سازد. هرچه فرآیند هدف‌گیری مکانیزه‌تر و مبتنی بر داده‌های پیچیده‌تر شود، اعمال استاندارد بالاتری مانند «قابلیت پیش‌بینی بهینه» برای تضمین رعایت حقوق بین‌الملل بشردوستانه، عملی‌تر و ضروری‌تر خواهد بود.

علاوه بر این، بعید است که دولت‌ها مخالفت اصولی جدی با استاندارد «قابلیت پیش‌بینی بهینه» داشته باشند، زیرا این استاندارد صرفاً در پی اعمال دقیق و کامل قانون موجود است. حقوق بین‌الملل بشردوستانه که همراه با اصول بنیادینش، اغلب به دلیل سازگاری با فناوری‌های نوین - مانند سلاح‌های هسته‌ای - مورد تحسین قرار گرفته است، باید به اندازه کافی

1. Reinforcement learning

انعطاف‌پذیر باشد تا بتواند استانداردهای خود را در مواجهه با تغییر پارادایم فناورانه ناشی از ظهور سلاح‌های یادگیرنده ارتقا دهد (Bhuta, Beck, & Geiss, 2016: 370). همان‌گونه که برخی پژوهشگران[1] استدلال می‌کنند، درحالی‌که شرایط ژئوپلیتیکی کنونی ممکن است مانع از بازنگری گسترده و رسمی قواعد موجود شود، حداقل یک تفسیر پیش‌رونده و پویا از قواعد موجود - که ویژگی‌های خاص سامانه‌های تسلیحاتی خودکار را در نظر می‌گیرد - نباید کنار گذاشته شود (Bhuta, Beck, & Geiss, 2016: 375)؛ بنابراین، استاندارد قابلیت پیش‌بینی بهینه به‌عنوان یکی از این‌گونه تفسیرهای پیش‌رونده و کارآمد از قوانین موجود پیشنهاد می‌شود.

3. تحلیل تطبیقی دیدگاه‌های مطرح در مواجهه با سلاح‌های خودکار یادگیرنده

در مواجهه با چالش‌های حقوقی سلاح‌های خودکار یادگیرنده، دکترین حقوق بین‌الملل بشردوستانه شاهد شکل‌گیری دیدگاه‌های متعددی است که می‌توان آن‌ها را در یک طیف از محافظه‌کاری حداکثری تا عمل‌گرایی مشروط قرار داد. دسته‌بندی و تحلیل تطبیقی این دیدگاه‌ها برای درک جایگاه نظری بحث و تبیین موضع این مقاله ضروری است.

3-1. دیدگاه مبتنی بر ممنوعیت یا محدودیت شدید (رویکرد احتیاطی- امنیتی)

این دیدگاه که مورد تأکید بسیاری از سازمان‌های غیردولتی و برخی حقوقدانان است، بر ویژگی‌های ذاتی سامانه‌های یادگیرنده - مانند

1. نهال بوتا، سوزان بک و رابین گایس: پژوهشگران حوزه حقوق بین‌الملل بشردوستانه و فناوری‌های نظامی نوین که به بررسی سامانه‌های تسلیحاتی خودمختار و چالش‌های حقوقی و اخلاقی ناشی از کاربرد هوش مصنوعی در میدان‌های نبرد می‌پردازند.

غیرقابل‌پیش‌بینی بودن، عدم شفافیت (تابلو باکس) و وابستگی به داده‌های آموزشی - تأکید می‌کند. استدلال محوری این گروه آن است که «مراقبت یک فرآیند ذهنی و انسانی است و نمی‌توان آن را به‌طور کامل به ماشینی واگذار کرد که عملکردی غیر شفاف یا غیرقابل‌پیش‌بینی دارد» (Schmitt, 2013: 48-49). از آنجا که این ویژگی‌ها، تحقق «کنترل انسانی معنادار» و «تعهد به مراقبت مستمر» را ناممکن یا بسیار مخدوش می‌سازند، نتیجه می‌گیرند که این سامانه‌ها ذاتاً با اصول بنیادین IHL ناسازگارند. برخی از صاحب‌نظران با استناد به «غیرقابل‌پیش‌بینی بودن ساختاری» این سامانه‌ها، خواستار «رویکردی احتیاطی» می‌شوند که ممکن است به محدودیت‌های شدید یا حتی ممنوعیت بینجامد (Bhuta et al., 2016: 372; Grut, 2013). این دیدگاه، ریسک گریز است و اولویت را بر پیشگیری کامل از نقض احتمالی حقوق بشردوستانه می‌گذارد.

2-3. دیدگاه اصلاح‌گرای کنترل انسانی (رویکرد الزام محور)

این دیدگاه می‌پذیرد که فناوری جدید، الگوهای سنتی کنترل را به چالش می‌کشد؛ اما به‌جای ممنوعیت، در پی بازتعریف و تقویت چارچوب‌های کنترلی موجود است. نمونه‌های شاخص این دیدگاه عبارت‌اند از:

- الگوی «کنترل انسانی معنادار در سطح تاکتیکی» (Roff & Moyes, 2016): که بر لزوم تأیید انسانی پیش از هر حمله تأکید دارد.
- الگوی «احتیاط پویا» (Margulies, 2017): که نظارت انسانی را ضروری ولی مشروط و قابل تفویض در شرایط خاص می‌داند.

محور مشترک این دیدگاه‌ها، اصرار بر حفظ نوعی از مداخله یا نظارت انسانی در حلقه عملیاتی است، هرچند شکل آن را انعطاف‌پذیر می‌دانند.

این دیدگاه همچنین بر «قابلیت توضیح» به‌عنوان پیش‌شرط کنترل معنادار تأکید دارد، زیرا «وقتی فرآیند تصمیم‌گیری یک سامانه نتواند توسط عامل انسانی تفسیر یا بازسازی شود، نسبت دادن مسئولیت، ارزیابی قانونی بودن یا مداخله معنادار غیرممکن می‌شود» (Mittelstadt et al., 2016: 9). این دیدگاه عموماً توسط دولت‌هایی که توسعه فناوری را دنبال می‌کنند اما نگران مشروعیت بین‌المللی هستند، مورد استناد قرار می‌گیرد.

3-3. دیدگاه مبتنی بر پیش‌بینی پذیری و مسئولیت‌پذیری پیشینی (رویکرد عملکردگرا)

این دیدگاه که توسط عالمانی مانند شولر (2017) صورت‌بندی شده، معتقد است معیار اصلی قانونی بودن نه لزوماً مداخله انسانی در لحظه، بلکه قابلیت پیش‌بینی معقول رفتار سامانه نسبت به اصول IHL است. شولر استدلال می‌کند که «الزام قانونی این نیست که هر عمل یک سلاح خودکار قابل پیش‌بینی باشد، بلکه این است که انطباق آن با حقوق بشردوستانه به‌طور معقول قابل پیش‌بینی باشد... اگر چنین پیش‌بینی معقولی از طریق برنامه‌نویسی، آزمون و محدودیت‌ها در طراحی عملیاتی قابل حصول باشد، آنگاه قانون مداخله انسانی در سطح تاکتیکی را به‌عنوان یک ضرورت مطلق طلب نمی‌کند» (Schuller, 2017: 408). این دیدگاه، منطقاً با تأکید بر «انتقال کانون مسئولیت‌پذیری به نقطه طراحی، آزمون و تأیید» همسو است (Cummings, 2017: 17). تمرکز این دیدگاه از کنترل عملیاتی به مسئولیت‌پذیری و تضمین کیفیت در مراحل پیش از به‌کارگیری منتقل می‌شود.

3-4. تحلیل تطبیقی و موضع‌گیری مقاله حاضر

مقاله حاضر ضمن پذیرش اهمیت چارچوب‌های پیشین، با انجام یک تحلیل تطبیقی، موضع خود را بر اساس قوت‌ها و ضعف‌های هر دیدگاه روشن می‌سازد.

در تقابل با دیدگاه اول (ممنوعیت): این مقاله استدلال می‌کند که رویکرد احتیاطی حداکثری، اگرچه نگرانی‌های اخلاقی مهمی را منعکس می‌کند، اما فرصت‌های فناوری را نادیده می‌گیرد. ظرفیت سامانه‌های یادگیرنده در پردازش داده‌های کلان و انجام محاسبات پیچیده - همان‌گونه که در نمونه‌هایی مانند AlphaGo Zero مشاهده شد - می‌تواند در شرایطی خاص به تصمیم‌گیری منطبق‌تر با اصول تفکیک و تناسب بینجامد (Silver & Hassabis, 2017)؛ بنابراین، ممنوعیت کامل به معنای از دست دادن یک ابزار بالقوه برای کاهش تلفات غیرنظامیان است.

در نقد دیدگاه دوم (الزام به کنترل عملیاتی): این مقاله با استناد به ماهیت سامانه‌های یادگیرنده استدلال می‌کند که در مواجهه با سیستم‌های پیچیده‌ای که سریع‌تر و با اطلاعات زمین‌های بیشتری از انسان تصمیم می‌گیرند، الزام به تأیید یا لغو انسانی در حلقه عملیاتی می‌تواند خود عاملی برای کاهش دقت و افزایش خطا باشد. این امر به دلیل «فاصله شناختی» بین اپراتور انسانی و ماشین و نیز پدیده «سوگیری خودکارسازی» رخ می‌دهد که در آن انسان ممکن است علی‌رغم شواهد، به سیستم بیش‌ازحد اعتماد کند یا در زمان نامناسب در آن دخالت کند (Matthias, 2004). در سطح کلان، مواجهه دولت‌ها با چنین فناوری‌های راهبردی، تابعی از محاسبات امنیتی و منافع ملی آن‌ها در نظام بین‌الملل است (اسمعیلی، ۱۴۰۳: ۶۱). ازاین‌رو، مدل‌های مبتنی بر کنترل عملیاتی برای این فناوری ناکارآمد هستند.

در توسعه و تکمیل دیدگاه سوم (پیش‌بینی پذیری): این مقاله، چارچوب شولر را جهت‌گیری درستی می‌داند که تمرکز را به جایگاه درست (مراحل پیشینی) معطوف می‌کند. با این حال، استدلال می‌کند که این چارچوب ناقص است. نقص

اصلی آن، عدم ارائه معیار مشخص و الزام‌آوری برای «میزان» پیش‌بینی پذیری لازم و غفلت از مدیریت ریسک ناشی از تعامل پیچیده انسان و ماشین و نیز خطر ذاتی سوگیری در داده‌های آموزشی است. استاندارد «پیش‌بینی معقول» شولر می‌تواند بسیار انعطاف‌پذیر تفسیر شود.

درنتیجه تحلیل فوق، گرایش غالب دکترینال در میانه دیدگاه دوم (اصلاح کنترل) و دیدگاه سوم (پیش‌بینی پذیری) در نوسان است. مقاله حاضر با پذیرش چارچوب «پیش‌بینی پذیری» به‌عنوان اساس، در پی تکمیل و تعمیق آن است. موضع نهایی این مقاله ارائه یک چارچوب عمل‌گرا و متوازن تحت عنوان «پیش‌بینی پذیری بهینه» است که در آن:

۱. معیار، نه صرفاً پیش‌بینی پذیری معقول، بلکه دستیابی به بالاترین سطح عملی از اطمینان نسبت به انطباق، با تکیه‌بر آزمون‌های جامع و متنوع است.
۲. کانون نظارت و مسئولیت‌پذیری انسانی به‌طور کامل به مراحل طراحی، آموزش، اعتبارسنجی و تصویب نهایی سامانه منتقل می‌شود.
۳. مدیریت ریسک‌های ویژه‌ای مانند سوگیری داده‌ها و تعامل انسان-ماشین به‌عنوان بخشی جدایی‌ناپذیر از این فرآیند نظارت پیشینی در نظر گرفته می‌شود.

این چارچوب از یکسو با تعهد به مراقبت مستمر سازگار است (زیرا حداکثر تلاش عملی برای اطمینان از رعایت قانون را طلب می‌کند) و از سوی دیگر، با بهره‌گیری از ظرفیت‌های فناوری، امکان تحقق عینی بهتر اصول تفکیک و تناسب و درنهایت کاهش تلفات غیرنظامیان را فراهم می‌آورد.

بخش ۳: تحلیل تطبیقی دیدگاه‌های دکترینی در مواجهه با سلاح‌های خودکار یادگیرنده

دیدگاه ممنوعیت‌محور

(رویکرد احتیاطی-امنیتی)

- نظریه‌پردازان: سازمان‌های حقوق بشری، گراوت (۲۰۱۲)، بوتا و همکاران (۲۰۱۶)
- تأکید بر غیرقابل پیش‌بینی بودن ذاتی سامانه‌های یادگیرنده
- مغایرت ذاتی با اصل «کنترل انسانی معنادار» در حقوق بشردوستانه
- پیشنهاد ممنوعیت کامل یا محدودیت‌های شدید
- اولویت دادن به پیشگیری کامل از نقض حقوق بشردوستانه

دیدگاه کنترل عملیاتی

(رویکرد الزام‌محور)

- نظریه‌پردازان: راف و موین (۲۰۱۶)، مارگولیس (۲۰۱۷)
- بازتعریف و تقویت چارچوب‌های کنترلی موجود
- اصرار بر مداخله یا نظارت انسانی در حلقه عملیاتی
- الگوی «کنترل انسانی معنادار در سطح تاکتیکی» (راف و موین)
- الگوی «احتیاط پویا» با امکان تفویض اختیار در شرایط خاص (مارگولیس)

دیدگاه پیش‌بینی‌پذیری

(رویکرد عملکردگرا)

- نظریه‌پرداز اصلی: شولر (۲۰۱۷)
- تغییر معیار از «مداخله» به «پیش‌بینی‌پذیری معقول» رفتار سامانه
- انتقال کانون کنترل به مراحل پیشینی (طراحی، آموزش، آزمون)
- تمرکز بر مسئولیت‌پذیری پیشینی و تضمین کیفیت
- امکان تفویض کامل اختیار در صورت پیش‌بینی‌پذیری کافی

چارچوب پیشنهادی مقاله

(پیش‌بینی‌پذیری بهینه و مراقبت پیشینی)

- ارائه شده در این مقاله
- توسعه نظریه شولر با الزام «پیش‌بینی‌پذیری بهینه» (بالاترین سطح عملی اطمینان)
- حکمرانی چرخه عمر سامانه با مدیریت یکپارچه ریسک
- توجه ویژه به چالش‌های سوگیری داده و تعامل انسان-ماشین
- سازگاری کامل با تعهد دولت‌ها به «مراقبت مستمر»

معیار ارزیابی	دیدگاه ممنوعیت‌محور	دیدگاه کنترل عملیاتی	دیدگاه پیش‌بینی‌پذیری	چارچوب مقاله حاضر
راهکار اصلی	ممنوعیت کامل توسعه و کاربرد	حفظ نظارت انسانی لحظه‌ای در عملیات	استانداردسازی و آزمون پیش از استقرار	حکمرانی یکپارچه چرخه عمر سامانه
نقاط قوت	سهولت حداکثری از اصول بشردوستانه	حفظ پیوند عُلّی انسان با تصمیم مرگبار	سازگاری با ماهیت فناوری‌های پیچیده	عملگرایی + مدیریت ریسک‌های نوظهور
نقاط ضعف	غفلت از پتانسیل کاهش تلفات غیرنظامی	ناسازگاری با سرعت و پیچیدگی سیستم‌ها	ابهام در تعریف استانداردهای «معقول»	پیچیدگی اجرایی و نیاز به همکاری بین‌رشته‌ای
جهت‌گیری در دکترین	نماینده دیدگاه‌های اقراطی احتیاطی	گرایش سنتی در حال افول	گرایش نوظهور و غالب در حال شکل‌گیری	نگه‌دارنده و توسعه‌دهنده گرایش غالب
نظریه‌پردازان شاخص	گراوت، بوتا و همکاران	راف و مویز، مارگولیس	شوار	مقاله حاضر

تحلیل تطبیقی و موضع‌گیری مقاله حاضر

مقاله با نقد نظام‌مند سه دیدگاه فوق، چارچوبی ترکیبی ارائه می‌دهد که از «پیش‌بینی‌پذیری» به عنوان پایه استفاده کرده، اما با افزودن الزام «هیئت‌سازی» و «مراقبت پیشینی»، کاستی‌های آن را مرتفع می‌سازد. این چارچوب ضمن حفظ تعهد دولت‌ها به مراقبت مستمر (ماده ۵۷ پروتکل الحاقی اول)، امکان بهره‌گیری مسئولانه از مزایای فناوری برای افزایش دقت در شناسایی اهداف و کاهش تلفات غیرنظامی را فراهم می‌آورد.

فرجام سخن

پژوهش حاضر با هدف واکاوی این پرسش کلیدی آغاز شد که نقش سلاح‌های خودکار مبتنی بر هوش مصنوعی و کنترل انسانی در حقوق بین‌الملل بشردوستانه، ذاتاً یک «چالش» بی‌مهار است یا می‌توانند به‌گونه‌ای سامان یابد که به «تعاملی» سازنده بین قابلیت‌های فناورانه و الزامات حقوقی تبدیل شود؟ یافته‌های این تحقیق نشان می‌دهد که رابطه مذکور، به‌تنهایی در هیچ‌یک از این دو قطب نمی‌گنجد. الگوی سنتی کنترل انسانی که بر مداخله عملیاتی مستقیم و تصویب پیشینی هر حمله تکیه دارد، نه‌تنها در بسیاری از سناریوهای آینده با ماهیت این فناوری ناسازگار است (و بنابراین خود به یک چالش تبدیل می‌شود)، بلکه ممکن است با ایجاد «سوگیری خودکارسازی» یا ایجاد تأخیرهای حیاتی، عملاً عاملی برای کاهش دقت و افزایش خطر برای غیرنظامیان باشد.

بررسی دیدگاه‌های دکترینی مؤید این است که رویکرد غالب در ادبیات حقوقی، دیگر ممنوعیت مطلق نیست، بلکه در حال گذار از کنترل

سختگیرانه عملیاتی به سمت چارچوب‌های نظارتی انعطاف‌پذیرتر است که بر مسئولیت‌پذیری، پیش‌بینی پذیری و تضمین کیفیت در مراحل پیش از به‌کارگیری متمرکزند. در این میان، نظریه‌هایی مانند «احتیاط پویا» (مارگولیس، ۲۰۱۷) و به‌ویژه «پیش‌بینی معقول» (شولر، ۲۰۱۷)، گام‌هایی روبه‌جلو محسوب می‌شوند، زیرا تمرکز را از «چگونگی مداخله انسان» به «چگونگی اطمینان از انطباق سامانه با قانون» منتقل می‌کنند.

با این حال، این مقاله استدلال می‌کند که حتی این چارچوب‌های پیشرفته نیز ناقص هستند. نقطه‌ضعف مشترک آن‌ها، عدم ارائه مکانیسم‌های عملی و الزام‌آور برای مدیریت ریسک‌های خاص سامانه‌های یادگیرنده، ازجمله (۱) وابستگی حیاتی و سوگیری بالقوه در داده‌های آموزشی، (۲) پیچیدگی و غیرقابل شفافیت ذاتی (جعبه سیاه) و (۳) تعامل پیش‌بینی‌ناپذیر انسان و ماشین است. غفلت از این چالش‌ها می‌تواند منجر به نقض سیستماتیک اصول تفکیک و تناسب شود، حتی اگر سامانه از نظر فنی «معقول» به نظر برسد.

در پاسخ به این خلأ، مقاله حاضر چارچوبی دوپایه تحت عنوان «پیش‌بینی پذیری بهینه مبتنی بر مراقبت پیشینی» را پیشنهاد می‌دهد. این چارچوب، تکلیف مراقبت دائمی را نه در لحظه شلیک که در مرحله طراحی، آموزش، آزمون و اعتبارسنجی نهایی سامانه جستجو می‌کند و بر دو اصل استوار است:

1) جایگزینی معیار «مداخله» با معیار «پیش‌بینی پذیری بهینه و قابل اعتبارسنجی»: در این مدل، قانونی بودن یک سامانه مشروط به دستیابی به بالاترین سطح عملی از اطمینان نسبت به انطباق رفتار آن با اصول حقوق بین‌الملل بشردوستانه است. این اطمینان نه صرفاً با آزمون‌های محدود، بلکه از طریق

شبیه‌سازی‌های گسترده، متنوع و شامل سناریوهای حاشیه‌ای و نیز ارزیابی مستمر کیفیت و بی‌طرفی داده‌های آموزشی باید حاصل شود. نقش حیاتی انسان، تصویب نهایی سامانه پس از اطمینان از احراز این استاندارد است، نه نظارت بر تکتک اقدامات آن.

2) انتقال کانون نظارت و مسئولیت‌پذیری به مرحله پیشینی: این انتقال مستلزم ایجاد چارچوب‌های حکمرانی و ممیزی فنی-حقوقی جدید در درون دولت‌ها و نهادهای نظامی است. در این چارچوب، متخصصان فنی، حقوق‌دانان و فرماندهان نظامی مشترکاً مسئول طراحی پروتکل‌های آزمون، تدوین معیارهای عملکردی شفاف و نظارت بر فرآیند یادگیری سامانه برای رفع سوگیری‌ها و اطمینان از تفسیرپذیری نسبی تصمیمات کلیدی هستند. این همان کنترل انسانی معنادار در عصر هوش مصنوعی است: خرد جمعی بشری در بالاترین سطح، معطوف به مهار و هدایت فناوری پیش از رهاسازی آن در میدان نبرد.

درنهایت، این چارچوب پیشنهادی نه درصدد تضعیف حقوق بین‌الملل بشردوستانه که در جهت تکامل کارکردی آن در مواجهه با یک تغییر پارادایم است. هدف غایی، همان «حفاظت از جمعیت غیرنظامیان» (پروتکل الحاقی اول، ماده ۵۷) باقی می‌ماند، اما ابزار تحقق آن توسعه می‌یابد. پذیرش این‌که در شرایطی، یک سامانه یادگیرنده می‌تواند با پردازش فرا بشری داده‌ها، تصمیمات دقیق‌تر و کم‌خطرتری نسبت به یک اپراتور انسانی تحت فشار اتخاذ کند، به معنای چشم‌پوشی از مسئولیت انسان نیست، بلکه به معنای تمرکز مسئولیت انسان در جایی است که بیشترین تأثیر را دارد: در خلق، آموزش و نظارت بر ماشینی که سپس می‌تواند به نمایندگی از او و در چارچوب ارزش‌های انسانیِ از پیش تعبیه‌شده، عمل کند.

بدین ترتیب، می‌توان از ظرفیت فناوری برای تقویت عینی اصول تفکیک و تناسب و کاهش رنج غیرنظامیان بهره برد، درحالی‌که کنترل نهایی و پاسخگویی در قبال آن، همواره در دایره اراده و خرد انسانی باقی می‌ماند.

منابع

اسمعیلی، مرتضی (۱۴۰۳). مناسبات قدرت و دکترین مسئولیت حمایت؛ مطالعه موردی مخاصمه ۲۰۲۳ حماس ـ اسرائیل، *مجله مطالعات حقوقی*، ۱۶ (۱)، ۴۳-۷۴.

پارسا، ناهید (۱۴۰۳). ناکارآمدی قوانین بیمه‌ای موجود در حوادث ناشی از استقلال هوش مصنوعی (مطالعه موردی خودروهای تمام خودران)، *مجله مطالعات حقوقی*، ۱۶ (۴)، ۷۵-۱۰۸.

علائی، صابر؛ حسین زاده، جواد (۱۴۰۱). واکاوی استقلال الگوریتم‌های جعبه سیاه در قراردادهای الگوریتمی و پیامدهای حقوقی آن، *مجله مطالعات حقوقی*، ۱۴ (۱)، ۲۵۱-۲۷۸.

DOI: 10.22099/jls.2022.40211.4340

References

af Jochnick, C., & Normand, R. (1994). The impact of technology on international humanitarian law. *International Review of the Red Cross*, 76(835), 1–30. https://doi.org/10.1017/S0020860400089648

Akerson, D. (2013). The illegality of offensive lethal autonomy. In D. Saxon (Ed.), *International humanitarian law and the changing technology of war* (pp. 65–87). Leiden: Brill Nijhoff.

Alaee, Saber & Hosseinzadeh, Javad (2022). Analysis of the Independence of Black Box Algorithms in Algorithmic Contracts and Its Legal Consequences. *Journal of Legal Studies*, 14(1), 251-278. DOI: 10.22099/jls.2022.40211.4340 [In Persian]

Article 36. (2013, April). Killer robots: UK government policy on fully autonomous weapons. London: Article 36. (Retrieved 10 Aban 1402) from: http://www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf

Bhuta, N., Beck, S., & Geiss, R. (2016). Present futures: Concluding reflections and open questions on autonomous weapons systems. In N. Bhuta, S. Beck, R. Geiss, H.-Y. Liu, & C. Kreß (Eds.), *Autonomous weapons systems: Law, ethics, policy* (pp. 347–375). Cambridge: Cambridge University Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Boulanin, V., & Verbruggen, M. (2017). *Mapping the development of autonomy in weapon systems*. Stockholm: Stockholm International Peace Research Institute (SIPRI). https://doi.org/10.13140/RG.2.2.22719.41127

Cummings, M. L. (2006). Automation and accountability in decision support system interface design. *Journal of Technology Studies*, 32(1), 23–31. https://www.jstor.org/stable/43604352

Curtis E. LeMay Center for Doctrine Development and Education. (2019). Volume 1: Basic doctrine. Montgomery: Curtis E. LeMay Center. (Retrieved 10 Aban 1402) from: https://www.doctrine.af.mil/Portals/61/documents/AFDP_3-0/3-0-AFDP-BASIC-DOCTRINE.pdf

de Spiegeleire, S., Maas, M., & Sweijs, T. (2017). *Artificial intelligence and the future of defense: Strategic implications for small- and medium-sized force providers*. The Hague: Stockholm International Peace Research Institute (SIPRI).

Deeks, A., Lubell, N., & Murray, D. (2019). Machine learning, artificial intelligence, and the use of force by states. *Journal of National Security Law & Policy*, 10(1), 1–34. https://jnslp.com/2019/12/02/machine-learning-artificial-intelligence-and-the-use-of-force-by-states/

Esmaeili, Morteza (2024). The Interplay of Power and the Responsibility to Protect Doctrine; A Case Study of the 2023 Hamas-Israel Conflict. *Journal of Legal Studies*, 16(1), 43-74. [In Persian]

Gill, T., Fleck, D., Boothby, W., & Vanheusden, A. (2014). Autonomous weapons and international humanitarian law: Challenges and perspectives. *Journal of Conflict and Security Law*, 19(1), 15–40. https://doi.org/10.1093/jcsl/kru001

Goose, S. (2014). *Autonomous weapons and international humanitarian law: A policy perspective*. Geneva: United Nations Institute for Disarmament Research (UNIDIR).

Gretton, A. (2017). *Legal and ethical issues in AI and autonomous weapons*. London: Routledge.

Grut, C. (2013). The challenge of autonomous lethal robotics to international humanitarian law. *Journal of Conflict and Security Law*, 18(1), 5–23. https://doi.org/10.1093/jcsl/krs026

Hassabis, D., & Silver, D. (2017, October 18). *AlphaGo Zero: Learning from scratch*. DeepMind Blog. (Retrieved 10 Aban 1402) from: https://deepmind.com/blog/alphago-zero-learning-scratch/

Henckaerts, J.-M., & Doswald-Beck, L. (2005). Customary international humanitarian law: A contribution to the understanding and respect for the rule of law in armed conflict. *International Review of the Red Cross*, 87(857), 51–80. https://doi.org/10.1017/S1816383100181130

Host, P. (2016, November 5*). Deep learning analytics develops DARPA deep machine learning prototype*. Defense Daily. (Retrieved 10 Aban 1402) from: https://www.defensedaily.com/

Human Rights Watch. (2015). *Mind the gap: The lack of accountability for killer robots*. New York: Human Rights Watch. (Retrieved 10 Aban 1402) from: https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots

International Committee of the Red Cross (ICRC). (2014). *Autonomous weapon systems: Technical, military, legal and humanitarian aspects*. Geneva: ICRC. (Retrieved 10 Aban 1402) from: https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014

Kania, E. B. (2017, June 8). *China's quest for an AI revolution in warfare. The Strategy Bridge*. (Retrieved 10 Aban 1402) from: https://thestrategybridge.org/the-bridge/2017/6/8/chinas-quest-for-an-ai-revolution-in-warfare

Keller, J. (2015, July 24). *DARPA TRACE program using advanced algorithms. Military Aerospace*. (Retrieved 10 Aban 1402) from: https://www.militaryaerospace.com/articles/2015/07/hpec-radar-target-recognition.html

Knight, W. (2017). *The dark secret at the heart of AI. MIT Technology Review*. (Retrieved 10 Aban 1402) from: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Lant, K. (2017, September 12). *China, Russia and the US are in an artificial intelligence arms race*. Futurism. (Retrieved 10 Aban 1402) from: https://futurism.com/china-russia-and-the-us-are-in-an-artificial-intelligence-arms-race

Margulies, P. (2017). Making autonomous weapons accountable: Command responsibility for computer-lethal force in armed conflicts. In J. D. Ohlin (Ed.), *Research handbook on remote warfare* (pp. 405–442). Cheltenham: Edward Elgar Publishing.

Marra, W. C., & McNeil, S. K. (2013). Understanding "the loop": Regulating the next generation of war machines. *Harvard Journal of Law & Public Policy*, 36(3), 1139–1186. https://heinonline.org/HOL/P?h=hein.journals/hjlpp36&i=1153

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

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). *The ethics of algorithms: Mapping the debate*. Big Data & Society, 3(2), 1–21. https://doi.org/10.1177/2053951716679679

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (2001). Aircrews and automation bias: The advantages of teamwork? *The International Journal*

of Aviation Psychology, 11(1), 1–14. https://doi.org/10.1207/S15327108IJAP1101_1

Neslage, K. (2015). Does "meaningful human control" have potential for the regulation of autonomous weapon systems? *National Security and Armed Conflict Law Review*, 6, 151–180. https://heinonline.org/HOL/P?h=hein.journals/nsecar6&i=155

Parsa, Nahid (2024). Inefficiency of Existing Insurance Laws in Accidents Caused by the Independence of Artificial Intelligence (A Case Study of Fully Self-Driving Cars). *Journal of Legal Studies*, 16(4), 75-108. [In Persian]

Pilloud, C., de Preux, J., Swinarski, C., & Zimmermann, B. (Eds.). (1987). *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949*. Geneva: International Committee of the Red Cross. https://doi.org/10.1163/9789004277285

Program on Humanitarian Policy and Conflict Resolution. (2010). *Autonomous weapons systems: Legal and ethical issues*. Cambridge: Harvard University.

Protective Edge. (2015). *Report on military operations in Gaza*. New York: United Nations Office for the Coordination of Humanitarian Affairs (OCHA).

Roff, H. M., & Moyes, R. (2016, April). "*Meaningful human control, artificial intelligence and autonomous weapons*": Briefing paper prepared for the informal meeting of experts on lethal autonomous weapons systems. London: Article 36. (Retrieved 10 Aban 1402) from: http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf

Scharre, P. (2016, February). *Autonomous weapons and operational risk. Washington, D.C.: Center for a New American Security (CNAS)*. (Retrieved 10 Aban 1402) from: https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, 29(2), 353–400. https://jolt.law.harvard.edu/assets/articlePDFs/v29/29HarvJLTech353.pdf

Schmitt, M. N. (2013). The use of autonomous weapons systems under international law. *Harvard National Security Journal*, 4(1), 20–55. https://harvardnsj.org/wp-content/uploads/sites/13/2013/01/Schmitt_Final.pdf

Schmitt, M. N. (2017). Autonomous weapon systems and international humanitarian law: A reconceptualization. *Harvard National Security*

Journal, 8, 1–12. https://harvardnsj.org/wp-content/uploads/sites/13/2017/06/Schmitt-Autonomous-Weapon-Systems-and-IHL.pdf

Schmitt, M. N. (Ed.). (2017). *Tallinn manual 2.0 on the international law applicable to cyber operations* (2nd ed.). Cambridge: Cambridge University Press.

Schuller, A. L. (2017). At the crossroads of control: The intersection of artificial intelligence in autonomous weapon systems with international humanitarian law. *Harvard National Security Journal*, 8(2), 379–435. https://harvardnsj.org/wp-content/uploads/sites/13/2017/11/Schuller-Final.pdf

Silver, D., & Hassabis, D. (2017). *AlphaGo Zero: Learning from scratch*. DeepMind Blog. (Retrieved 10 Aban 1402) from: https://deepmind.com/blog/alphago-zero-learning-scratch/

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018, December 6). AlphaZero: Shedding new light on chess, shogi, and Go. DeepMind Blog. (Retrieved 10 Aban 1402) from: https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

Simonite, T. (2017, July 19). AI could revolutionize war as much as nukes. WIRED. (Retrieved 10 Aban 1402) from: https://www.wired.com/story/ai-could-revolutionize-war-as-much-as-nukes/

Smith, B. W. (2016). Lawyers and engineers should speak the same robot language. In R. Calo, A. M. Froomkin, & I. Kerr (Eds.), *Robot law* (pp. 98–117). Cheltenham: Edward Elgar Publishing.

Stensmo, M., & Sejnowski, T. J. (1996). Neural networks, autonomous agents, and the law of armed conflict. *Journal of Law, Information and Science*, 6(2), 45-60. https://doi.org/10.5778/JLIS.1996.6.stensmo.45

Switzerland, Permanent Mission of Switzerland to the United Nations Office and other international organizations in Geneva. (2016, March 30). Towards a "compliance based" approach to LAWS. Geneva: Permanent Mission of Switzerland. (Retrieved 10 Aban 1402) from: https://documents.unoda.org/wp-content/uploads/2016/04/20160330_Switzerland.pdf

The Economist. (2018, March 15). *America v China—The battle for digital supremacy*. The Economist. (Retrieved 10 Aban 1402) from: https://www.economist.com/leaders/2018/03/15/the-battle-for-digital-supremacy

Trapp, K. (2013). *Great resources mean great responsibility: A framework of analysis for assessing compliance with API obligations in the information age.* In D. Saxon (Ed.), International humanitarian law and the changing technology of war (pp. 159–180). Leiden: Brill Nijhoff.

Turner, J. (2018). Robot rules: Regulating artificial intelligence. London: Palgrave Macmillan.

Wagner, M. (2014). The dehumanization of international humanitarian law: Legal, ethical, and political implications of autonomous weapon systems. *Vanderbilt Journal of Transnational Law*, 47(5), 1371–1424. https://scholarship.law.vanderbilt.edu/vjtl/vol47/iss5/3/

Treaties & Judicial

Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (CCW), Protocol III on Prohibitions or Restrictions on the Use of Incendiary Weapons, October 10, 1980, 1342 U.N.T.S. 171.

Hague Convention (IV) respecting the Laws and Customs of War on Land and its annex: Regulations concerning the Laws and Customs of War on Land, October 18, 1907, 205 Consol. T.S. 277.

International Criminal Tribunal for the Former Yugoslavia. (2000, November 16). Prosecutor v. Delalić et al. (Case No. IT-96-21-T), Judgment.

Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226 (July 8).

Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Additional Protocol I), June 8, 1977, 1125 U.N.T.S. 3.

Rome Statute of the International Criminal Court, July 17, 1998, 2187 U.N.T.S. 90.