첨단기술과 형사법 국제세미나
International Seminar on
Advanced Technologies and Criminal Justice

첨단기술과 형사사법의 미래

Advanced Technologies and the Future of Criminal Justice

KICJ 한국형사·법무정책연구원
Korean Institute of Criminology and Justice

SEOUL AI HUB

개회사

여러분 안녕하십니까. 한국형사·법무정책연구원 원장 정웅석입니다. 〈첨단기술과 형사사법의 미래〉라는 주제로 개최되는 국제세미나에 함께해 주신 모든 분들께 진심으로 감사드립니다.

먼저 바쁘신 일정 속에서도 이 자리를 빛내 주신 해외 전문가 여러분께 깊이 감사드립니다. 독일 범죄예방대회의 에리히 막스(Erich Marks) 대표님, 프랑스 보비니 검찰청의 에릭 마테(Eric Mathais) 검사장님, 그리고 중국 서북정법대의 첸징춘(Chen Jingchun) 교수님께서 본 세미나를 위해 귀한 시간을 내주셨습니다. 형사사법 분야에서 국제적으로 중요한 역할을 수행해 오신 세 분의 통찰은 오늘의 논의를 더욱 풍부하고 깊이 있게 만들어 줄 것이라 확신합니다.

아울러 국내 학계와 실무계를 대표해 사회와 토론을 맡아 주시는 여러 전문가 여러분께도 깊은 감사의 말씀을 드립니다. 김성돈 교수님, 한명관 고문님, 한상훈 교수님께서 각 세션의 사회를 맡아 주셨으며, 조기영 교수님을 비롯한 많은 전문가께서 첨단기술이 제기하는 형사법적 도전과 그 대응 방향에 대해 깊이 있는 논의를 이끌어 주실 예정입니다. 진심으로 감사드립니다.

특히, 오늘 패널토론에는 형사법 연구를 선도하는 5대 학회의 회장단께서 함께해 주십니다. 급변하는 기술 환경 속에서 형사사법 체계가 나아갈 방향을 학문적·정책적으로 통합하여 제시할 수 있는 매우 뜻깊은 자리이며, 저 역시 큰 기대를 가지고 있습니다. 각 학회가 축적해 온 경험과 전문성이 오늘의 논의를 통해 더욱 심화되고, 인공지능 시대에 국가 형사정책의 수립에 실질적으로 기여하는 성과로 이어지리라 믿습니다.

한국형사·법무정책연구원은 인공지능, 바이오 기술, 디지털 제조, 양자컴퓨팅 등 첨단기술이 형사·법무 분야에 미치는 영향을 조망하고, 인권과 안전이 조화되는 미래 지향적 형사사법 체계를 구축하기 위한 연구를 지속해 왔습니다. 오늘 세미나는 이러한 연구 성과를 공유하며, 각국의 다양한 경험과 관점을 모아 새로운 정책적 방향을 모색하기 위한 뜻깊은 자리입니다.

첨단기술은 우리에게 새로운 가능성을 제공하는 동시에, 기존의 형사사법 체계가 충분히 예상하지 못했던 도전 과제를 함께 제기하고 있습니다. 오늘의 논의가 기술혁신 시대에 부합하는 형사사법의 새로운 기준을 제시하고, 국제적 협력을 한층 강화하는 데 의미 있는 기여를 할 수 있기를 기대합니다.

끝으로, 이번 행사를 공동으로 준비해 주신 서울 AI 허브 관계자 여러분께도 깊이 감사드립니다. 여러분의 협력과 지원이 있었기에 오늘의 자리가 성공적으로 마련될 수 있었습니다. 참석해 주신 모든 분들께 다시 한 번 감사드리며, 오늘 세미나가 첨단기술 시대 형사사법 정책 발전을 위한 소중한 밑거름이 되기를 바랍니다.

감사합니다.

2025년 11월

한국형사·법무정책연구원

원장 **정 웅 석**

Welcome Speech

It is my great pleasure to deliver these congratulatory remarks at this meaningful event.
My name is Chanjin Park, Director of the Seoul AI Hub.

It is an honor to offer these remarks at a seminar organized by the Korean Institute of Criminology and Justice. Today's gathering takes place at a moment of profound technological transition, one that is reshaping not only our industries but also the broader environment in which law and public safety operate.

In earlier years, artificial intelligence was limited to narrow tasks such as identifying images or detecting defects in manufacturing. But today's AI has evolved far beyond those early capabilities. Modern AI systems can understand the semantic context of images and videos, perform step-by-step reasoning, and gathers and synthesizes external information when needed, and carries out actions based on planning and decision-making.

At present, AI operates primarily through integration with IT systems, but within the next two to three years, it is expected to be embedded in robots and autonomous platforms, extending its capabilities into the physical world. This marks the beginning of the era of agentic AI—technology that not only analyzes information but also takes action. This shift presents new opportunities for innovation, while also raising important questions regarding safety, accountability, and governance—questions that the legal community is uniquely positioned to help address.

Over the past two years, intelligence has been concentrated largely in horizontal AI services such as chatbots, translation, and search. However, we are now entering the early stages of a transition where AI increasingly moves vertically into core industries such as finance, manufacturing, mobility, healthcare, and public safety. The global competition that once centered on enhancing foundation models —such as ChatGPT and Gemini—has evolved into a race to apply AI effectively and responsibly across real-world domains.

As AI becomes more deeply integrated into society, the landscape of crime, security threats, and social risks is also beginning to change. We are already observing early signs of AI misuse in areas such as elections, examinations, and authentication systems. These developments underscore the importance of collaboration among the criminal justice community, policymakers, and technology experts. The legal field, with its long tradition of evaluating responsibility, fairness, and public interest, plays a critical role in guiding how AI should—and should not—be used.

In this rapidly evolving environment, legal and regulatory frameworks are becoming increasingly vital. They must ensure that AI serves human dignity and societal well-being, while remaining flexible enough to support responsible innovation. Technology and law must evolve in dialogue, each informing and strengthening the other.

For the criminal justice community, AI presents both new challenges and important new tools. When thoughtfully applied, AI can support evidence-based decision-making, enhance public safety, detect emerging risks, and contribute to a more resilient and trustworthy justice system. Realizing these benefits requires informed discussion—exactly the kind of dialogue that this seminar is designed to foster.

This is why today's event is so meaningful. By bringing together experts in criminal law, public policy, and emerging technologies, this seminar provides an invaluable platform for examining how AI will shape the future of justice and how each field can contribute to ensuring that these changes benefit society.

I hope today's discussions offer deep insight and inspire continued collaboration between the legal and technology communities as we navigate this transformative era. Thank you.

November, 2025

Park, Chanjin

Director, Seoul AI Hub

첨단기술과 형사사법의 미래
Advanced Technologies and the Future of Criminal Justice

목차

첨단기술과 형사법 국제세미나

1. 주제: **첨단기술과 형사사법의 미래**
2. 일자: **2025년 11월 25일(화)**
3. 시간: **13:00 - 18:00**
4. 장소: **서울 AI Playground(한국교총 1층)**
5. 주최: **한국형사·법무정책연구원 / 서울 AI 허브**
6. 세부 일정

시간		내 용
12:30 - 13:00	30	**등 록**
13:00 - 13:20	20	• 사 회: **윤지영**(한국형사·법무정책연구원 선임연구위원) • 개회사: **정웅석**(한국형사·법무정책연구원 원장) • 환영사: **박찬진**(서울 AI 허브 센터장) • 기념촬영 ※ 동시통역 제공
13:20 - 13:30	10	**휴 식**
13:30 - 15:10	100	**제1주제** **AI 시대의 범죄예방과 데이터 보안** • 사 회: **김성돈**(성균관대 법학전문대학원 교수, 한국형사법학회 고문) • 발 표: **Erich Marks**(독일 범죄예방대회 대표) **Chen Jingchun**(중국 서북정법대 교수) • 토 론: **조기영**(전북대 법학전문대학원 교수) **안수길**(명지대 법학과 교수) **고명수**(서울대 법학전문대학원 교수) **조형찬**(한국형사·법무정책연구원 부연구위원)
15:10 - 15:30	20	**휴 식**
15:30 - 17:10	100	**제2주제** **첨단기술의 발전과 범죄, 그리고 형사사법의 대응** • 사 회: **한명관**(법무법인 바른 변호사, 제4차산업혁명융합법학회 고문) • 발 표: **Eric Mathais**(프랑스 보비니 검찰청 검사장) **윤지영**(한국형사·법무정책연구원 선임연구위원) • 토 론: **이근우**(가천대 법학과 교수) **김대원**(인하대 법학전문대학원 초빙교수) **이원상**(조선대 법학과 교수) **류부곤**(경찰대 법학과 교수)
17:10 - 17:20	10	**휴 식**
17:20 - 17:50	30	**패널토론** **AI와 형사사법의 미래: 쟁점과 과제** • 사 회: **한상훈**(연세대 법학전문대학원 교수, 한국형사법학회 고문) • 토 론: **황태정**(경기대 경찰행정학과 교수, 한국형사법학회 회장) **김성룡**(경북대 법학전문대학원 교수, 한국형사소송법학회 회장) **최호진**(단국대 법학과 교수, 한국비교형사법학회 회장) **이경렬**(성균관대 법학전문대학원 교수, 한국피해자학회 회장) **김한균**(한국형사·법무정책연구원 선임연구위원, 한국형사정책학회 회장)
17:50 - 18:00	10	**폐회사**

International Seminar on Advanced Technologies and Criminal Justice

1. Theme: **Advanced Technologies and the Future of Criminal Justice**
2. Date: **Tuesday, November 25, 2025**
3. Time: **13:00 – 18:00**
4. Venue: **Seoul AI Playground(KFTA 1F)**
5. Organizer: **Korean Institute of Criminology and Justice/Seoul AI Hub**
6. Detailed schedule

Time		Note	
12:30 – 13:00	30	**Registration**	
13:00 – 13:20	20	• Moderator: **Yun, Jee-Young**(Senior Research Fellow, Korean Institute of Criminology and Justice) • Opening Remarks: **Jeong, Woong Seok**(President, Korean Institute of Criminology and Justice) • Congratulatory Remarks: **Park, Chanjin**(Managing Director, Seoul AI Hub Center) • Commemorative Photo ※ simultaneous interpretation provided	
13:20 – 13:30	10	**Break**	
13:30 – 15:10	100	**Session 1**	**Crime Prevention and Data Security in the AI Era**
		• Moderator:	**Kim, Seong Don**(Professor, Sungkyunkwan University Law School)
		• Presentation:	**Erich Marks**(Managing Director, German Prevention Congress) **Chen Jingchun**(Professor, Northwest University of Political Science and law)
		• Discussion:	**Cho, Giyeong**(Professor, Jeonbuk National University Law School) **An, Sugil**(Professor, Department of Law, Myongji University) **Ko, Myoung-su**(Professor, Seoul National University Law School) **Jo, Hyoung Chan**(Research Fellow, Korean Institute of Criminology and Justice)
15:10 – 15:30	20	**Break**	
15:30 – 17:10	100	**Session 2**	**Technological Advancements, Crime, and Response of Criminal Justice**
		• Moderator:	**Hahn, Myung Kwan**(Partner, Barun Law LLC/Advisor)
		• Presentation:	**Eric Mathais**(State prosecutor, Bobigny Judicial Court) **Yun, Jee-Young**(Senior Research Fellow, Korean Institute of Criminology and Justice)
		• Discussion:	**Lee, Keun-woo**(Professor, Department of law, Gachon University) **Kim, Dae Won**(Visiting Professor, Inha University Law School) **Lee, Won-Sang**(Professor, Department of law, Chosun University) **Ryu, Bu-Gon**(Professor, Department of law, Korean National Police University)
17:10 – 17:20	10	**Break**	
17:20 – 17:50	30	**Panel Discussion**	**AI and the Future of Criminal Justice: Issues and Challenges**
		• Moderator:	**Han, Sang Hoon**(Professor, Law School of Yonsei University)
		• Discussion:	**Hwang, Tae-jeong**(Professor, Department of Police Administration, Kyonggi University) **Kim, Sung-Ryong**(Professor, Kyungpook National University Law School) **Choi, Ho-Jin**(Professor, Department of Law, Dankook University) **Lee, Kyung-Lyul**(Professor,Sungkyunkwan University Law School) **Kim, Han-Kyun**(Senior Research Fellow, Korean Institute of Criminology and Justice)
17:50 – 18:00	10	**Closing Remarks**	

Session I

Crime Prevention and Data Security in the AI Era

AI 시대의 범죄예방과 데이터 보안

Moderator: Kim, Seong Don

Professor, Sungkyunkwan University Law School

Advisor, Korean Criminal Law Association

사회: 김성돈

성균관대 법학전문대학원 교수

한국형사법학회 고문

Session I

Ki in der Prävention

예방 분야에 있어서의 인공지능

Erich Marks

Managing Director,

German Prevention Congress

독일 범죄예방대회 대표

International Seminar on Advanced Technologies and Criminal Justice

Korean Institute of Criminology and Justice
Seoul 11/2025

Erich Marks: “KI in der Prävention”

Gliederung

1 Deutscher Präventionstag und “KI in der Prävention”

2 zu einigen allgemeinen KI-Aspekten

3 ausgewählte Aspekte und Einzelthemen

4 Aufgaben und Folgen für die Prävention

1

Deutscher Präventionstag (DPT)

&

“KI in der Prävention”

31. Deutscher Präventionstag

- Der Deutsche Präventionstag (**DPT**) ist der weltweit größte Jahreskongress zur Kriminalprävention sowie angrenzender Präventionsbereiche.
- Der Kongress wendet sich an Verantwortungsträger der Prävention in Kommunen, bei der Polizei, im Gesundheitswesen, in der Jugendhilfe, in der Justiz, in den Religionsgemeinschaften, im Bildungsbereich, in Vereinen und Verbänden sowie an Politiker und Wissenschaftler.
- Der 31. Deutsche Präventionskongress (**DPT**) findet am 13. und 14. April 2026 im Congress Centrum Hannover statt.
- Das Schwerpunktthema dieses Kongresses lautet “KI in der Prävention”. Hierzu wird ein umfangreiches wissenschaftliches Gutachten erstellt, dass im Februar 2026 veröffentlicht wird.

Zentrale Fragen des Schwerpunktthemas

- Welche Herausforderungen bringt KI im Kontext von Kriminalität und Sicherheit, aber auch im gesamtgesellschaftlichen Miteinander mit sich?
- Welche tiefgreifenden Veränderungen gehen mit ihrem Einsatz einher – und wer ist davon in welcher Weise betroffen?
- Wie lässt sich KI gezielt und verantwortungsvoll für die Präventionsarbeit nutzen?
- Dabei geht es nicht nur um technologische Potenziale, sondern auch um die ethische und praktische Frage, wie ein bewusster, reflektierter Umgang mit KI in der Prävention gelingen kann.

Wissenschaftliche Begleitschrift

- Im Vorfeld des Kongresses wird eine wissenschaftliche Begleitschrift erstellt, in der das Schwerpunktthema aus verschiedenen wissenschaftlichen Perspektiven aufbereitet wird. Zusammengefasst wird diese in einem Kurzvideo. Die Gesamtkoordination erfolgt durch Prof. Dr. Gina Rosa Wollinger.
- Begleitschrift und Video werden im Frühjahr 2026 auf der Webseite www.praeventionstag.de veröffentlicht.
- Ein kurzer Einführungstext fasst die Thematik zusammen.

Wissenschaftliche Expertisen (1)

- **Einleitung** (Prof. Dr. Gina Rosa Wollinger)
  Inhalt: Spannungsfeld von KI und Prävention, weite der Anwendungsmöglichkeiten, kriminologische Bezüge, Kosten/Schattenseiten von KI, Vorstellung der einzelnen Beiträge
- **Geleitwort (**Dr. M. Fübi, A. Schneider)
  Inhalte: Cybersicherheit, KI und Cybersicherheit, Doppelrolle der KI (Angriff aber auch Prävention), Wie KI sicher einsetzen? Lässt sich KI prüfen?
- **KI als Bias-Fall? (**Prof. Dr. Alke Martens)
  Inhalte: Was ist KI?, Was ist ein Bias?, Datenbias und Algorithmenbias, Bias in KI – Ursachen und Folgen
- **Rechtliche Herausforderungen von Innovation bis Anwendung**
  (Prof. Dr. Sebastian Golla)
  Inhalte: KI als Instrument zur Kriminalprävention, Eingriffsrechtliche und datenschutzrechtliche Einordnung, Neue Herausforderungen durch die KI-Verordnung der EU, Rechtliche Hürden der KI-Innovation

Wissenschaftliche Expertisen (2)

- **Künstliche neuronale Netze im Strafverfahren – Zwischen Chancen und Risiken (**Alina Borowy)
  Inhalte: Einsatz im Strafverfahren, Assistenzleistung und eigene Wissensgenerierung, Mögliche Einsatzszenarien (Data Mining, Gesichtserkennung, Videoüberwachung KI), Erprobungen in der Praxis, Risiken und Bedenken bei der Nutzung, Bias und Blackbox-Effekt, Falsche Treffer und Trefferquoten, Schwere der Grundrechtseingriffe
- **Predictive Policing und Kriminalprävention – Chancen und Grenzen algorithmischer Prognosen**
  (Dr. Simon Egbert)
  Inhalte: Aufkommen von KI in der polizeilichen Präventionsarbeit, Definition und Funktionsweise von Predictive Policing, Prävention als Ziel: von situativer Kriminalprävention bis zu algorithmischer Prognose, Abgrenzung zu klassischen Formen der (präventiven) Polizeiarbeit, Predictive Policing als soziotechnische Interaktion: Die Relevanz der Umsetzung von Kriminalitätsprognosen für erfolgreiche Prävention, Empirische Einsatzfelder und Praktiken, Predictive Policing als soziotechnische Interaktion: Die Relevanz der Umsetzung von Kriminalitätsprognosen für erfolgreiche Prävention, Predictive Policing und „repressive“ Prävention

Wissenschaftliche Expertisen (3)

- **Die Beiträge von KI an extremistischer Kommunikation: Ist (generative) KI bereits teilnehmende Einheit sozialer Interaktion?**
  (Dr. Christian Büscher, Prof. Dr. Isabel Kusche, Tim Röller, Alexandros Gazos)
  Inhalte: Technologiemonitoring im Kontext von Radikalisierung und Extremismus, Nutzung von KI durch extremistische Akteure, Prävention des Missbrauchs von KI-Anwendungen, KI-Anwendungen als Ressource für die Prävention von Radikalisierung und Extremismus
- **DeTox und BoTox: Projekte zur Unterstützung der Bekämpfung von Hasskriminalität im Netz**
  (Prof. Dr. Melanie Sigel & Florian Meyer)
  Inhalte: Detektion von Toxizität und Aggressionen in Postings und Kommentaren im Netz, Bot- und Kontexterkennung im Umfeld von Hasskommentaren

Auswahl einiger KI-Vorträge im April 2026 (1)

- Einsatz von KI für Risikoanalysen
- Notrufkompetenz im Kindesalter durch KI-Lernsysteme
- KI als Brücke zum Hilfesystem bei häuslicher Gewalt
- Generative KI zur Prävention von Menschenfeindlichkeit
- Online-offline referrals to P/CVE services in the age of AI
- Digitale Radikalisierung: TikTok, KI und Prävention
- Hate Speech Prevention & KI
- KI-unterstützte kommunale Sicherheitsanalysen

Auswahl einiger KI-Vorträge im April 2026 (2)

- KI als Radikalisierungs- und Extremisierungsbeschleuniger?
- KI bei richterlichen Kriminalprognosen
- Civic Resilience AI - Generative KI vs. Diskriminierung
- KI im Kulturgüterschutz
- Radikalisierungsprävention mit KI in der Schule
- KI und Jugendschutz
- Cybersicherheit und KI in Unternehmen - Der Faktor Mensch
- Online-Betrug und KI: Daten & Maßnahmen

Auswahl einiger KI-Vorträge im April 2026 (3)

- KI-gestützte Prävention von Menschenhandel
- KI-gestützter Gefährdungskoeffizient für den Zufahrtsschutz
- KI als neue Herausforderung für die Opferhilfe
- KI und der Verlust des Menschlichen?
- KI in der Straffälligenhilfe
- KI, Macht & Missbrauch: Deepfakes, Chatbots & Algorithmen
- KI als Frühwarnsystem in der Prävention
- KI, Deepfakes und Opferwerdung im digitalen Raum

Zu einigen Erfahrungen des DPT mit KI

- KI-unterstützte Kongresseröffnung des 28. DPT (2023) in Mannheim
- "Der kleiner Präventionsprinz"
- Avatar-Einsatz bei DPT-TV und fremdsprachigen Vorträgen
- Die täglichen Präventions-News (TPN) des Deutschen Präventionstages informieren seit 2011 in deutscher sowie englischer Sprache über Aktuelles aus den Bereichen Präventionspraxis, Präventions-forschung und Präventionspolitik. Seit Juli 2025 erscheinen wöchentlich Nachrichten zum Thema „KI in der Prävention".

Praxisumfrage von DPT und ZHAW

Anfang 2026 wird der Deutsche Präventionstag (DPT) in Kooperation mit Prof. Dr. Dirk Baier, Direktor des Instituts für Delinquenz und Kriminalprävention der Zürcher Hochschule für angewandte Wissenschaften (ZHAW, Swiss) eine Umfrage zu Erfahrungen, Planungen und Sichtweisen auf Künstliche Intelligenz durchführen. Die Umfrage richtet sich an Institutionen und Experten der Kriminalprävention in Deutschland, Österreich und der Schweiz. Erste Ergebnisse werden dann im Rahmen des 31. Deutschen Präventionstages im April 2026 in Hannover vorgestellt.

2

Zu einigen

allgemeinen Aspekten

der Künstlichen Intelligenz

Bias in der KI

- Bundesamt für Sicherheit in der Informationstechnik (BSI): Bias-Whitepaper
- Mittelstand-Digital Zentrum Zukunftskultur: Künstliche Intelligenz und Bias
- SAP: Bias in der künstlichen Intelligenz
- Universität Zürich: KI bewertet Texte neutral – bis sie die Quelle kennt

Lernplattformen für KI

- KI-Campus beim Stifterverband:
  Lernplattformen für Künstliche Intelligenz
- Wikipedia versus Grokipedia
- 360learning:
  KI-gestützte Lernplattform für kollaboratives Lernen

Regulierungen der KI

- Europäisches Parlament (EP):
  Europäische Verordnung zur künstlichen Intelligenz
- Wikipedia: Regulierung von künstlicher Intelligenz
- Globaler Aufruf zu roten Linien der KI
- D64: Code of Conduct Demokratische KI
- Bertelsmann Stiftung: "Simplifying" European AI Regulation: An Evidence-based White Paper

KI & Ethik

- Klicksafe:
  10 Gebote der KI-Ethik
- Klaus Tschira Stiftung:
  KI als moralischer Dialogpartner
- Europäische Union (EU):
  Ethikleitlinien für vertrauens-würdige KI
- Wie wir KI wirklich demokratisieren (Publix)
  Prof. Dr. Annette Zimmermann

In Deutschland beliebte KI-Podcast-Angebote

- Deutschlandfunk: "KI verstehen"
- ARD: "Der KI-Podcast"
- Frankfurter Allgemeine Zeitung (FAZ): "Künstliche Intelligenz"
- fobizz.com: "Kreide. KI. Klartext."
- Zentrum für vertrauenswürdige Künstliche Intelligenz (ZVKI): "Trust Issues"
- Deutsches Forschungszentrum für Künstliche Inteligenz (DFKI) & Radio Berlin Brandenburg (rbb): "KI – und jetzt"
- Heise Online: "KI-Update" und "Deep Mindes"
- Charité & Stifterverband: "Dr. med. KI"
- KI-Campus-Community: "KI kapiert"
- Schwarz Digits: "Tech, KI und Schmetterlinge"

3

Zu einigen ausgewählten

Aspekten und Einzelthemen

von

“KI in der Prävention”

Vorbemerkung

Künstliche Intelligenz (KI) verändert tiefgreifend nahezu alle Lebensbereiche – auch jene, die mit Kriminalität, Sicherheit und sozialem Zusammenleben verbunden sind. Ihr Einsatz birgt enorme Chancen, zugleich aber erhebliche Herausforderungen.

KI in der Polizeiarbeit

- Innovationslabor von Europol
  Leitfaden
- Wissenschaftliches Side-Event auf dem Europäischen Polizeikongress 2025: „Künstliche Intelligenz in der Polizeiarbeit“
- Bitkom-Positionspapier:
  „KI in der Polizei – Einsatzpotentiale und Lösungsansätze zur Implementierung“

KI in der Strafjustiz

- Goldman Sachs
  Große Entwicklungspotentiale auch für die Justiz
- Vals Legal AI Report (VLAIR), KI teils besser als Anwältinnen und Anwälte (VLAIR+)
- Max-Planck-Institut:
  Algorithmisches Profiling und automatisierte Entscheidungsfindung in der Strafjustiz
- Anwalt.de: KI und Strafrecht

KI im Gesundheitswesen

- Hasso Plattner Institut (HPI):
  Mit KI gesund bleiben
- Institut für Gesundheitsgestaltung (nuvio):
  KI in Public Health & Gesundheitsförderung
- Deutsches Krebsforschungs-zentrum (DKFZ):
  Langfristige Prognosen von Krankheitsrisiken

KI & Kriminalität

- ZAHW, Prof. Dr. Dirk Baier
  Künstliche Intelligenz und Kriminalität
- ProPK:
  Künstliche Intelligenz im Alltag und in der Kriminalitätsbekämpfung
- EVOLUCE:
  Kann künstliche Intelligenz Verbrechen vorhersagen?

KI im Bildungsbereich

- Deutsches Institut für Erwachsenenbildung (DIE):
  KI für lebenslanges Lernen
- Digitale Lernwoche 2025 der UNESCO
- Universität Bochum (RUB):
  Bedeutung der KI-Verordnung für Bildungseinrichtungen
- KI-Monitor 2025 (Stifterverband)
- Telekom: Trendmonitor KI in der Bildung

KI in der Schule

- Robert Bosch Stiftung:
  Deutsches Schulbarometer
- Rahmenprogramm empirische Bildungsforschung:
  Künstliche Intelligenz in der Schule. Eine Handreichung zum Stand in Wissenschaft und Praxis
- Telli:
  der KI-Chatbot für die Schule

KI und urbane Sicherheit

- Bundesamt für Bauwesen und Raumordnung (BBSR)
  Künstliche Intelligenz in smarten Städten und Regionen
- Deutscher Städtetag (DST):
  Mit KI und Geoinformationen: Wie Urbane Digitale Zwillinge die Stadtentwicklung revolutionieren
- Friedrich Naumann Stiftung:
  Stadt und KI

Integration von KI in Präventionsprogramme

- Charité Berlin, Prof. Dr. Dr. K.M. Beier
  „Kein Täter werden“
- 35. Niedersächsische Suchtkonferenz:
  Künstliche Intelligenz in der Suchthilfe und Suchtprävention
- Grüne Liste Prävention

1. Herausforderungen im Kontext von Kriminalität und Sicherheit

KI-Systeme können sowohl Werkzeug als auch Tatmittel sein:

- Neue Kriminalitätsformen entstehen, etwa durch automatisierte Cyberangriffe, Deepfakes oder KI-gestützte Betrugsstrategien.
- Manipulation und Desinformation werden durch generative KI erleichtert und gefährden gesellschaftliches Vertrauen.
- Datenschutz und Überwachung: KI-gestützte Analyse großer Datenmengen kann einerseits Sicherheitsbehörden unterstützen, andererseits aber tief in die Privatsphäre eingreifen und Grundrechte gefährden.
- Die Herausforderung besteht darin, Sicherheit zu stärken, ohne Freiheit und Vertrauen zu verlieren.

2. Gesellschaftliche Veränderungen und Betroffenheit

Der Einsatz von KI führt zu strukturellen und kulturellen Veränderungen:

- Polizei und Justiz müssen neue Kompetenzen aufbauen und rechtliche Rahmenbedingungen anpassen.
- Bürgerinnen und Bürger erleben KI-basierte Systeme zunehmend im Alltag (z. B. Gesichtserkennung, Social Scoring, algorithmische Risikobewertung).
- Soziale Ungleichheiten können sich verschärfen, wenn algorithmische Systeme Vorurteile reproduzieren („Bias“) oder bestimmte Gruppen benachteiligen.
- Damit wird KI zu einem gesellschaftspolitischen Thema, das Fragen nach Transparenz, Kontrolle und Verantwortung aufwirft.

3. Chancen und verantwortungsvolle Nutzung in der Präventionsarbeit

Richtig eingesetzt, kann KI auch präventiv wirken – etwa durch:

- Früherkennung von Risiken, z. B. bei Cyberkriminalität, Hate Speech oder Radikalisierungstendenzen in sozialen Medien.
- Analyse sozialer Netzwerke, um potenzielle Gefährdungslagen zu erkennen, ohne einzelne Menschen pauschal zu verdächtigen.
- Unterstützung der Bildungs- und Aufklärungsarbeit, indem KI Lernprozesse personalisiert oder Informationskampagnen zielgerichtet gestaltet.
- Zugleich verlangt Prävention im digitalen Zeitalter ein ethisch reflektiertes Handeln:
- KI darf kein Ersatz für menschliches Urteilsvermögen sein, sondern ein Werkzeug, das verantwortungsvoll genutzt wird.
- Es braucht Transparenz, Nachvollziehbarkeit und Partizipation, um Vertrauen zu schaffen.
- Bildung und Medienkompetenz sind entscheidend, damit Fachkräfte und Bürger gleichermaßen KI-kompetent und kritisch handeln können.

Fazit

Die zentrale Herausforderung besteht darin, KI als Werkzeug gesellschaftlicher Verantwortung zu begreifen – nicht als bloße Technologie.

Nur durch interdisziplinäre Zusammenarbeit von Technologie, Ethik, Recht, Pädagogik und Zivilgesellschaft kann es gelingen, die Potenziale von KI für Prävention und Sicherheit zu nutzen, ohne die Werte einer demokratischen Gesellschaft zu gefährden.

4

Zentrale Anregungen zur weiteren Bearbeitung des Themenfeldes „KI in der Prävention“

1. Ethische und normative Grundlagen

- Welche Werte und Prinzipien (z. B. Menschenwürde, Datenschutz, Nichtdiskriminierung, Transparenz) müssen den Einsatz von KI in der Prävention leiten?
- Wie lässt sich eine „Ethik der Verantwortung“ konkret in die Entwicklung, Anwendung und Bewertung von KI-Systemen integrieren?
- Wie kann gesellschaftliche Kontrolle (z. B. Ethikräte, zivilgesellschaftliche Beteiligung) sichergestellt werden?

2. Praktische Einsatzfelder und Nutzenpotenziale

- In welchen präventiven Handlungsfeldern (z. B. Kriminalprävention, Gewaltprävention, Extremismusprävention, Suchtprävention, Jugendarbeit, Verkehrssicherheit) kann KI sinnvoll eingesetzt werden?
- Wie können KI-basierte Frühwarnsysteme gestaltet werden, ohne zu stigmatisieren oder zu überwachungslastig zu sein?
- Welche best-practice-Beispiele zeigen bereits, dass KI die Prävention unterstützen kann (z. B. Mustererkennung in Cybercrime, Analyse von Fake News, Früherkennung von Gefährdungslagen)?

3. Bildung, Kompetenzaufbau und Qualifizierung

- Welche digitalen und ethischen Kompetenzen benötigen Fachkräfte in Prävention, Polizei, Sozialarbeit und Bildung, um KI verantwortungsvoll nutzen zu können?
- Wie kann KI-Kompetenz in Aus- und Weiterbildung integriert werden (z. B. durch Fortbildungen, Planspiele, Reflexionsmodule)?
- Welche Rolle spielen Bildung und Aufklärung auch gegenüber der Bevölkerung, um Ängste abzubauen und ein reflektiertes Verständnis von KI zu fördern?

4. Forschung, Evaluation und Evidenzbildung

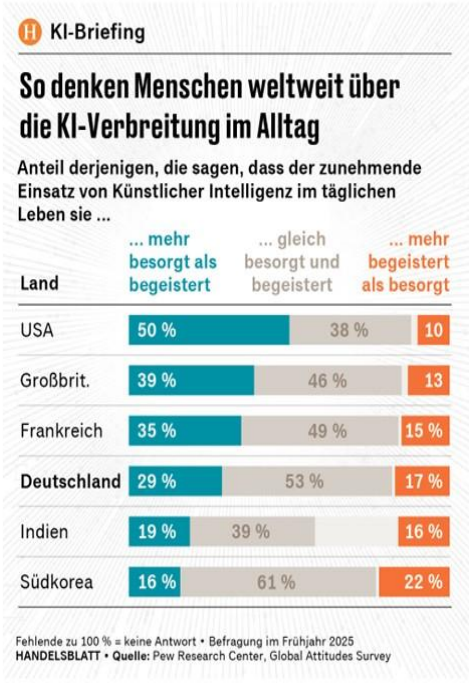
- Welche methodischen Ansätze eignen sich, um die Wirksamkeit und Fairness von KI-gestützten Präventionsmaßnahmen zu bewerten?
- Wie kann transdisziplinäre Forschung (Technik, Sozialwissenschaft, Kriminologie, Ethik) gefördert werden?
- Welche Daten und Indikatoren sind erforderlich, um KI-gestützte Präventionsstrategien evidenzbasiert weiterzuentwickeln – und wie lässt sich dies datenschutzkonform realisieren?

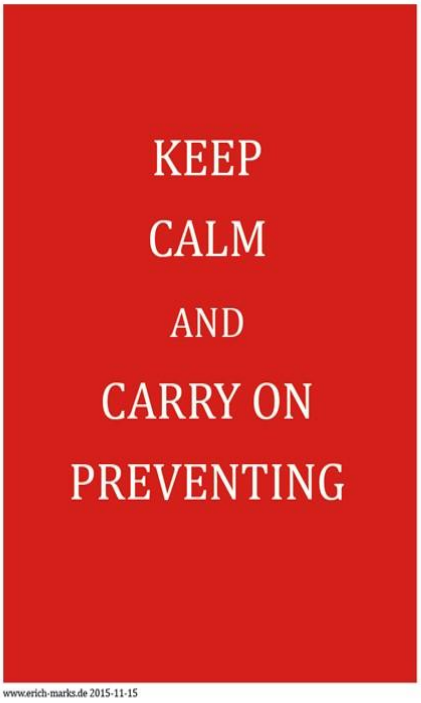
5. Rechtliche und politische Rahmenbedingungen

- Wie müssen gesetzliche Grundlagen (z. B. Datenschutz, algorithmische Entscheidungsfindung, Haftungsfragen) angepasst werden, um verantwortungsvollen KI-Einsatz zu ermöglichen?
- Welche politischen Steuerungsinstrumente (Förderprogramme, Standards, Zertifizierungen) sind geeignet, um Vertrauen und Qualität zu sichern?
- Wie kann internationale Kooperation im Bereich KI und Prävention gestaltet werden, um globale Entwicklungen zu berücksichtigen?

6. Gesellschaftliche Reflexion und Beteiligung

- Wie können Bürgerinnen und Bürger in Diskussionen über Chancen und Risiken von KI eingebunden werden?
- Welche Formate fördern einen öffentlichen Diskurs über KI, Sicherheit und Prävention – z. B. Dialogforen, Bürgerlabore oder partizipative Projekte?
- Wie kann eine Kultur der Achtsamkeit und kritischen Reflexion im Umgang mit KI entstehen, die sowohl Innovation als auch Verantwortung betont?





Die Bearbeitung des Themenfeldes „KI in der Prävention“ erfordert ein dauerhaftes Zusammenspiel von Technikentwicklung, Ethik, Bildung, Forschung, Politik und Praxis.
Ziel sollte es sein, Leitlinien und Modelle zu entwickeln, die den verantwortungsvollen Einsatz von KI fördern und gleichzeitig die Resilienz der Gesellschaft gegenüber Missbrauch und Fehlentwicklungen stärken.

첨단기술과 형사법
국제세미나

한국형사•법무정책연구원
서울 2025. 11. 25.

Erich Marks: “예방 분야에 있어서의 인공지능”

개 요

1 독일범죄예방대회(Deutscher Präventionstag)와
“예방 분야에 있어서의 인공지능(KI in der Prävention)”

2 일반적인 인공지능 관점에 대하여

3 선별된 관점 및 개별 주제

4 예방을 위한 과제와 결과

1

독일범죄예방대회(DPT)

&

“예방 분야에 있어서의 인공지능(KI in der Prävention)”

제31회 독일범죄예방대회

- 독일범죄예방대회(Deutsche Präventionstag, DPT)는 범죄예방 및 관련 예방 분야를 다루는 세계 최대 규모의 연례회의
- 이 대회는 지방자치단체, 경찰, 보건의료, 청소년 복지, 사법부, 종교 단체, 교육 분야, 협회 및 단체의 예방 책임자뿐만 아니라 정치인과 과학자들을 대상으로 한다.
- 제31회 독일범죄예방대회(DPT)는 2026년 4월 13일과 14일 하노버 컨벤션 센터에서 개최된다.
- 이번 대회의 주요 주제는 “예방 분야에 있어서의 인공지능(AI)”이다. 이에 대해 포괄적인 과학적 보고서가 작성되어 2026년 2월에 발표될 예정이다.

중점 주제의 핵심 질문

- 범죄 및 보안의 맥락에서, 그리고 사회 전체의 공존 측면에서 인공 지능이 가져오는 도전 과제는 무엇인가?
- 그 활용과 함께 발생하는 근본적인 변화는 무엇이며, 누가 어떤 방식으로 영향을 받는가?
- 예방 활동에 AI를 효과적이고 책임감 있게 활용하는 방법은 무엇인가?
- 이는 단순히 기술적 잠재력뿐만 아니라, 예방 분야에서 AI를 의식적이고 성찰적으로 활용하는 방법에 대한 윤리적·실용적 문제이기도 하다.

학술 부속보고서

- 학술대회에 앞서 주요 주제를 다양한 학문적 관점에서 다루는 학술 부속보고서가 제작되고, 그 내용은 짧은 동영상으로 요약 제공된다. 전체 총괄은 Gina Rosa Wollinger 교수가 담당하고 있다.
- 부록과 동영상은 2026년 봄에 웹사이트 www.praeventionstag.de에 공개될 예정이다.
- 주제를 요약한 간단한 소개글이 포함된다.

과학적 전문성 (1)

- **서론**(Gina Rosa Wollinger 교수)

  내용: 인공지능과 예방의 긴장 관계, 적용 가능성의 폭, 범죄학적 연관성, 인공지능의 비용/단점, 개별 기고문 소개

- **서문** (M. Fübi 박사, A. Schneider)

  내용: 사이버 보안, AI와 사이버 보안, AI의 이중 역할(공격이자 예방), AI를 안전하게 활용하는 방법은 무엇인가? AI를 검증할 수 있는가?

- **AI는 편향 사례인가?** (Alke Martens 교수)

  내용: AI란 무엇인가?, 편향이란 무엇인가?, 데이터 편향과 알고리즘 편향, AI 내 편향–원인과 결과

- **혁신에서 적용까지의 법적 과제** (Sebastian Golla 교수)

  내용: 범죄 예방 도구로서의 AI, 개입권 및 개인정보 보호법상 분류, EU AI 규정으로 인한 새로운 도전 과제, AI 혁신의 법적 장애물

과학적 전문성 (2)

- **형사 절차에서의 인공 신경망 – 기회와 위험 사이** (Alina Borowy)

  내용: 형사 절차에서의 활용, 보조 기능 및 자체 지식 생성, 가능한 활용 시나리오(데이터 마이닝, 얼굴 인식, 영상 감시 AI), 실제 적용 사례, 사용 시 위험 및 우려 사항, 편향성과 블랙박스 효과, 오탐지 및 적중률, 기본권 침해의 심각성

- **예측적 치안 및 범죄 예방 – 알고리즘 예측의 기회와 한계 (**Dr. Simon Egbert)
  내용: 경찰 예방 업무에서의 AI 등장, 예측형 치안의 정의와 작동 방식, 목표로서의 예방: 상황별 범죄 예방에서 알고리즘 예측까지, 전통적인 (예방적) 경찰 업무와의 구분, 사회기술적 상호작용으로서의 예측형 치안: 성공적인 예방을 위한 범죄 예측 구현의 중요성, 경험적 적용 분야 및 실천, 예측형 치안 활동으로서의 사회기술적 상호작용: 성공적인 예방을 위한 범죄 예측 구현의 중요성, 예측형 치안 활동과 "억제적" 예방

과학적 전문성 (3)

- **인공지능이 극단주의적 의사소통에 기여하는 바: (생성형) 인공지능은 이미 사회적 상호작용의 참여 주체인가?**
  (Christian Büscher 박사, Isabel Kusche 교수, Tim Röller, Alexandros Gazos)
  내용: 급진화와 극단주의 맥락에서의 기술 모니터링, 극단주의 행위자들의 AI 활용, AI 애플리케이션 오용 방지, 급진화와 극단주의 방지를 위한 자원으로서의 AI 애플리케이션
- **DeTox와 BoTox: 온라인 증오 범죄 퇴치를 지원하는 프로젝트**
  (Melanie Sigel 교수 & Florian Meyer)
  내용: 온라인 게시물 및 댓글에서 독성과 공격성 탐지, 증오 댓글 환경에서의 봇(Bot)및 컨텍스트 인식

2026년 4월 AI 강연 일부 선정(1)

- 위험 분석을 위한 AI 활용
- AI 학습 시스템을 통한 아동기의 응급 상황 대처 능력
- 가정폭력 지원 시스템을 위한 가교 역할로서의 AI
- 인간 혐오 예방을 위한 생성형 AI
- AI 시대의 P/CVE 서비스에 대한 온라인-오프라인 연계
- 디지털 급진화: 틱톡, AI 및 예방
- 혐오 발언 예방 및 AI
- AI-지원 지역 사회 안전 분석

2026년 4월 AI 강연 일부 선정(2)

- AI가 급진화와 극단화를 가속화하는 요인인가?
- 사법적 범죄 예측에 활용되는 AI
- 시민 회복탄력성 AI - 생성형 AI vs. 차별
- 문화재 보호를 위한 AI
- 학교에서 AI를 활용한 급진화 예방
- AI와 청소년 보호
- 기업 내 사이버 보안과 AI - 인간 요인
- 온라인 사기 및 AI: 데이터 및 대책

2026년 4월 AI 강연 일부 선정(3)

- 인공지능 기반 인신매매 예방
- 접근 통제를 위한 AI-기반 위험 계수
- 피해자 지원에 대한 새로운 도전으로서의 AI
- AI와 인간성의 상실?
- 범죄자 지원에서의 AI
- AI, 권력 및 남용: 딥페이크, 챗봇 및 알고리즘
- 예방을 위한 조기 경보 시스템으로서의 AI
- 디지털 공간에서의 인공지능, 딥페이크 및 피해자화

독일범죄예방대회의 AI 활용 경험 일부

- AI 지원으로 열린 제28회 독일범죄예방대회(2023) 만하임 대회 개막식
- "작은 예방의 왕자"
- DPT-TV 및 외국어 강연에서의 아바타 활용
- 독일범죄예방대회(DPT)의 일간 예방 뉴스(TPN)는 2011년부터 독일어와 영어로 예방 실무, 예방 연구 및 예방 정책 분야의 최신 소식을 전하고 있다. 2025년 7월부터 "예방 분야에 있어서의 AI"를 주제로 한 뉴스가 매주 발행되고 있다.

독일범죄예방대회(DPT)와 취리히 응용과학대학교(ZHAW)의 실무 설문조사

2026년 초, 독일범죄예방대회(DPT)는 취리히 응용과학대학교(ZHAW, 스위스) 범죄 및 범죄 예방 연구소 소장인 Dirk Baier 교수와 협력하여 인공지능에 대한 경험, 계획 및 관점에 대한 설문조사를 실시할 예정이다.

이 설문조사는 독일, 오스트리아, 스위스의 범죄 예방 기관 및 전문가들을 대상으로 한다. 첫 번째 결과는 2026년 4월 하노버에서 열리는 제31회 독일범죄예방대회 행사에서 발표될 예정이다.

2

인공지능의 몇 가지 일반적인 측면에 관하여

인공지능(AI)의 편향성

- 독일 연방정보기술보안청(BSI):
  편향 백서
- 미래 문화를 위한 중소기업 디지털 센터:
  인공 지능과 편향
- SAP:
  인공 지능 내 편향
- 취리히 대학교(Universität Zürich): AI는 텍스트를 중립적으로 평가한다, 출처를 알기 전까지는.

인공지능 학습 플랫폼

- 스티프터페어반트(Stifterverband)의 AI 캠퍼스:
  인공 지능 학습 플랫폼
- 위키피디아(Wikipedia) 對 그로키피디아(Grokipedia)
- 360learning:
  협업 학습을 위한 AI 기반 학습 플랫폼

인공지능 규제

- 유럽 의회(EP):
  유럽 인공지능 규정
- 위키백과(Wikipedia): 인공 지능 규제
- 인공지능의 레드라인에 대한 글로벌 호소
- D64: 민주적 인공지능 행동 강령
- 베르텔스만 재단(Bertelsmann Stiftung): 유럽 AI 규제 간소화: 증거 기반 백서

인공지능(AI)과 윤리

- 클릭세이프(Klicksafe):
  AI 윤리의 10계명

- 클라우스 치라 재단(Klaus Tschira Stiftung):
  도덕적 대화 상대로서의 AI

- 유럽 연합(EU):
  신뢰할 수 있는 AI를 위한 윤리 지침

- AI를 진정으로 민주화하는 방법(Publix)
  Annette Zimmermann 교수

독일에서 인기 있는 AI 팟캐스트 서비스

- 독일 방송(Deutschlandfunk): "AI 이해하기"
- ARD: "AI 팟캐스트"
- 프랑크푸르터 알게마이네 차이퉁(FAZ): "인공 지능"
- fobizz.com: "분필. AI. 명료한 텍스트."
- 신뢰할 수 있는 인공지능 센터(ZVKI): "신뢰 문제"
- 독일 인공지능 연구 센터(DFKI) & 베를린 브란덴부르크 라디오(rbb): "AI – 그리고 지금"
- 하이제 온라인(Heise Online): "AI 업데이트" 및 "딥 마인드"
- 샤리테 & 스티프터반드(Charité & Stifterverband): "의학박사 AI"
- AI 캠퍼스 커뮤니티: "AI 이해하기"
- 슈바르츠 디지츠(Schwarz Digits): "테크, AI 그리고 나비"

3

“예방 분야에 있어서의 인공지능 (KI in der Prävention)”에 관한 주요 논점과 개별 주제들

서문

인공지능(AI)은 범죄, 안전, 사회적 공존과 관련된 분야를 포함해 거의 모든 삶의 영역을 근본적으로 변화시키고 있다. 인공지능의 활용은 엄청난 기회를 제공하지만 동시에 상당한 도전 과제도 안고 있다.

경찰 업무에서의 인공지능

- 유로폴 혁신 연구소(Innovationslabor von Europol)
  가이드라인
- 2025년 유럽 경찰 회의 과학 부대행사: "경찰 업무에서의 인공지능"
- 비트콤 입장문(Bitkom-Positionspapier):
  "경찰 업무에서의 AI – 활용 가능성과 구현을 위한 해결 방안"

형사 사법에서의 인공지능

- 골드만 삭스(Goldman Sachs)
  사법 분야에서도 큰 발전 가능성이 있음
- 발스 법률 AI 보고서(VLAIR), AI가 변호사보다 더 나은 경우도 있음 (VLAIR+)
- 막스 플랑크 연구소 (Max-Planck-Institut):
  형사 사법에서의 알고리즘 프로파일링 및 자동화된 의사 결정
- Anwalt.de: AI와 형법

의료 분야의 인공지능

- 하소 플래트너 연구소(Hasso Plattner Institut, HPI):
  인공지능으로 건강 유지하기

- 건강 설계 연구소(Institut für Gesundheitsgestaltung(nuvio)):
  공공 보건 및 건강 증진 분야의 인공지능

- 독일 암 연구 센터(Deutsches Krebsforschungs-zentrum, DKFZ):
  질병 위험의 장기적 예측

인공지능(AI)과 범죄

- ZAHW, Dirk Baier 교수
  인공 지능과 범죄

- ProPK:
  일상생활과 범죄 퇴치에서의 인공지능

- EVOLUCE:
  인공지능이 범죄를 예측할 수 있을까?

교육 분야의 인공지능

- 독일 성인교육 연구소(Deutsches Institut für Erwachsenenbildung, DIE):
  평생 학습을 위한 인공지능

- 유네스코 디지털 학습 주간 2025(Digitale Lernwoche 2025 der UNESCO)

- 보훔 대학교(Universität Bochum, RUB):
  교육 기관을 위한 AI 규정의 중요성

- AI 모니터 2025(Stifterverband)

- 텔레콤(Telekom): 교육 분야 AI 트렌드 모니터

학교에서의 인공지능

- 로버트 보쉬 재단(Robert Bosch Stiftung):
  독일 학교 바로미터

- 교육 연구 프레임워크 프로그램:
  학교에서의 인공지능. 과학 및 실무 현황에 관한 안내서

- Telli:
  학교용 AI 챗봇

인공지능과 도시 안전

- 독일 연방건축공간연구원(BBSR)
스마트 도시 및 지역에서의 인공지능

- 독일 도시협의회(Deutscher Städtetag, DST):
AI와 지리정보를 통해: 도시 디지털 트윈이 도시 개발을 혁신하는 방법

- 프리드리히 나우만 재단(Friedrich Naumann Stiftung):
도시와 인공지능

인공지능(AI)의 예방 프로그램 통합

- 베를린 샤리테 병원(Charité Berlin), K.M. Beier 교수
"가해자가 되지 않기"

- 제35회 니더작센 중독 컨퍼런스:
중독 지원 및 중독 예방에 있어서의 인공 지능

- 예방을 위한 그린 리스트

1. 범죄와 안전의 맥락에서 직면한 과제

AI 시스템은 도구이자 범죄 수단이 될 수 있다:

- 자동화된 사이버 공격, 딥페이크, AI 기반 사기 전략 등 새로운 형태의 범죄가 등장하고 있다.
- 생성형 AI로 인해 조작과 허위 정보 유포가 용이해져 사회적 신뢰를 위협한다.
- 개인정보 보호와 감시: AI 기반의 대량 데이터 분석은 한편으로는 보안 당국을 지원할 수 있지만, 다른 한편으로는 사생활을 심하게 침해하고 기본권을 위협할 수 있다.
- 과제는 자유와 신뢰를 잃지 않으면서 보안을 강화하는 것이다.

2. 사회적 변화 및 영향

인공지능의 활용은 구조적·문화적 변화를 초래한다:

- 경찰과 사법부는 새로운 역량을 구축하고 법적 프레임워크를 조정해야 한다.
- 시민들은 일상생활에서 AI 기반 시스템(예: 얼굴 인식, 사회적 점수 매기기(social scoring), 알고리즘 위험 평가)을 점점 더 많이 경험하게 된다.
- 알고리즘 시스템이 편견("바이어스")을 재생산하거나 특정 집단을 불리하게 대우할 경우 사회적 불평등이 심화될 수 있다.
- 이로 인해 AI는 투명성, 통제, 책임에 대한 질문을 제기하는 사회정치적 주제로 부상하고 있다.

3. 예방 활동에서의 기회와 책임 있는 활용

적절하게 활용될 경우, AI는 다음과 같은 방식으로 예방적 효과를 발휘할 수 있다:

- 사이버 범죄, 증오 발언 또는 소셜 미디어에서의 급진화 경향과 같은 위험의 조기 발견.
- 개별 개인을 무분별하게 의심하지 않으면서 잠재적 위험 상황을 파악하기 위한 소셜네트워크 분석
- AI가 학습 과정을 개인화하거나 정보 캠페인을 목표 지향적으로 설계함으로써 교육 및 계몽 활동 지원
- 동시에 디지털 시대의 예방은 윤리적 성찰을 바탕으로 한 행동을 요구한다
- AI는 인간의 판단력을 대체해서는 안 되며, 책임감 있게 사용되는 도구여야 한다.
- 신뢰를 구축하기 위해서는 투명성, 추적 가능성 및 참여가 필요하다.
- 전문가와 시민 모두가 AI 역량을 갖추고 비판적으로 행동할 수 있도록 하는 데 있어서 교육과 미디어 리터러시가 결정적으로 작용한다.

결론

핵심 과제는 AI를 단순한 기술이 아닌 사회적 책임의 도구로 인식하는 것이다.

기술, 윤리, 법, 교육학 및 시민사회 사이의 학제 간 협력을 통해서만 AI의 잠재력을 예방과 안전에 활용하면서도 민주적 사회의 가치를 훼손하지 않을 수 있다.

4

"예방 분야에 있어서의 인공지능"

주제에 대한

향후 논의를 위한

핵심 제언

1. 윤리적 및 규범적 기초

- 예방 분야에서 AI 활용을 이끌어야 할 가치와 원칙(예: 인간의 존엄성, 개인정보 보호, 차별 금지, 투명성)은 무엇인가?
- "책임의 윤리"를 AI 시스템의 개발, 적용 및 평가에 구체적으로 어떻게 통합할 수 있을까?
- 사회적 통제(예: 윤리위원회, 시민사회 참여)는 어떻게 보장할 수 있을까?

2. 실용적 적용 분야 및 활용 가능성

- 어떤 예방적 활동 분야(예: 범죄 예방, 폭력 예방, 극단주의 예방, 중독 예방, 청소년 복지, 교통 안전)에서 AI를 효과적으로 활용할 수 있을까?
- 낙인 찍기나 과도한 감시 없이 AI 기반 조기 경보 시스템을 어떻게 설계할 수 있을까?
- 이미 AI가 예방을 지원할 수 있음을 보여주는 모범 사례는 무엇인가?(예: 사이버 범죄의 패턴 인식, 가짜 뉴스 분석, 위험 상황의 조기 발견)

3. 교육, 역량 강화 및 자격 취득

- 예방, 경찰, 사회복지 및 교육 분야의 전문가들이 AI를 책임감 있게 활용하기 위해 필요한 디지털 및 윤리적 역량은 무엇인가?
- 교육 및 연수 과정에 AI 역량을 어떻게 통합할 수 있을까?

  (예: 연수, 시뮬레이션 게임, 성찰 모듈)
- 인구 전체에 대한 교육과 계몽은 두려움을 해소하고 AI에 대한 성찰적 이해를 촉진하는 데 어떤 역할을 하는가?

4. 연구, 평가 및 증거 구축

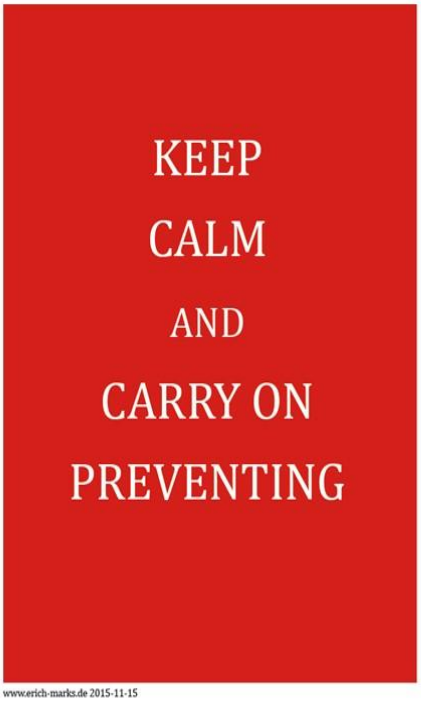
- AI 기반 예방 조치의 효과성과 공정성을 평가하기 위해 어떤 방법론적 접근이 적합한가?
- 기술, 사회과학, 범죄학, 윤리학 등 다양한 학문 분야를 아우르는 초학제적 연구를 어떻게 촉진할 수 있을까?
- AI 기반 예방 전략을 증거 기반 방식으로 발전시키기 위해 필요한 데이터와 지표는 무엇이며, 이를 개인정보 보호 규정에 부합하도록 구현하는 방법은 무엇인가?

5. 법적 및 정책적 프레임워크

- 책임 있는 AI 사용을 가능하게 하기 위해 법적 기반(예: 데이터 보호, 알고리즘 기반 의사 결정, 책임 문제)은 어떻게 조정되어야 하는가?
- 신뢰와 품질을 보장하기 위해 어떤 정책적 통제 수단(지원 프로그램, 표준, 인증)이 적합한가?
- 글로벌 발전을 고려하기 위해 AI 및 예방 분야에서의 국제 협력을 어떻게 구성할 수 있을까?

6. 사회적 성찰 및 참여

- 시민들은 AI의 기회와 위험에 대한 논의에 어떻게 참여할 수 있을까?
- AI, 안전 및 예방에 대한 공개적 논의를 촉진하는 형식은 무엇인가?(예: 대화 포럼, 시민 연구소 또는 참여형 프로젝트)
- 혁신과 책임을 모두 강조하는 AI 활용에 대한 주의 깊은 태도와 비판적 성찰의 문화를 어떻게 조성할 수 있을까?



"예방 분야에 있어서의 인공지능"이라는 주제를 다루기 위해서는 기술 개발, 윤리, 교육, 연구, 정책 및 실무 간의 지속적인 협력이 필요하다.
목표는 AI의 책임 있는 사용을 촉진하고 동시에 오용 및 잘못된 발전에 대한 사회의 회복탄력성을 강화하는 지침과 모델을 개발하는 것이어야 한다.

Session Ⅰ

大模型算法数据安全保护的刑法应对

(대규모 모델(大模型) 알고리즘의 데이터 안전 보호에 대한 형법적 대응)

Chen Jingchun(陈京春)

Professor,
Northwest University of
Political Science and law
중국 서북정법대학 교수

大模型算法数据安全保护的刑法应对

西北政法大学　陈京春

2025年11月

法治信仰 中国立场 国际视野 平民情怀

一、大模型数据安全风险对刑事法的挑战

（一）大规模性与法益的确定

1.大模型算法对个人和社会的影响具有大规模性。

2.涉大模型算法违法行为对集体法益的侵害具有现实性。

3.对于集体法益的强调需要谨防法益稀薄化的风险。

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW
嚴謹·求實·文明·公正

一、大模型数据安全风险对刑事法的挑战

（二）全过程性与危险的识别

1.大模型算法的研发带来了人为的安全风险。

2.大模型算法的部署使社会安全风险现实化。

3.大模型算法服务的使用行为亦存在法益侵害的可能。

法治信仰 中国立场 国际视野 平民情怀

一、大模型数据安全风险对刑事法的挑战

（三）不确定性与规制的策略

1.大模型算法的安全风险未并在研发阶段就能够被充分认知。

2.大模型算法的安全风险在部署阶段不断显现和发展变化。

3.大模型算法使用端的行为加剧了安全风险的不确定性。

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 嚴謹·求實·文明·公正

一、大模型数据安全风险对刑事法的挑战

（四）不可解释性与因果关系

1.大模型算法具有一定程度上的不可解释性。

2.大模型算法的不可解释性给因果关系的认定带来困难。

3.大模型算法的可解释性是可信人工智能亟需解决的课题。

法治信仰 中国立场 国际视野 平民情怀

一、大模型数据安全风险对刑事法的挑战

（五）人机交互性与刑事归责

1.大模型算法产品及其服务具有人机互动的显著特征。

2.基于信赖原则，善意的使用者通常情况下不应被追究刑事责任。

3.对于“恶意”使用行为，在何种情况下需要进行追责需要具体考量。

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 嚴謹·求實·文明·公正

二、刑法在大模型数据安全治理中的定位

（一）技术治理发挥着主导作用

1.对于大模型算法需要敏捷治理，及时做出技术回应。

2.技术治理在大模型数据安全治理中发挥着主导作用。

3.需要将法律规范的内容落实到技术（行为）规范层面。

法治信仰 中国立场 国际视野 平民情怀

二、刑法在大模型数据安全治理中的定位

（二）人工智能法居于法律体系的核心

1.人工智能法是行政法的核心。

2.人工智能法也是算法（数据处理）的安全法。

3.人工智能法为探寻刑法介入的边界和尺度提供了实验性的参照。

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 严谨·求實·文明·公正

二、刑法在大模型数据安全治理中的定位

（三）刑法的介入回应需坚守谦抑原则

1.传统网络犯罪的立法模式与司法适用逻辑产生了罅隙。

2.刑法的介入及其刑事归责应当保持极为审慎的态度。

3.是否转变刑事立法和刑事司法的模式需要深入探讨。

法治信仰 中国立场 国际视野 平民情怀

三、大模型数据安全保护的刑法归责理论

（一）多元主体间合理的危险分配

1.合规的研发者创设了被允许的风险。

2.部署者成为大模型算法危险防控的主体。

3.更大的危险来自基于非法目的的研发、利用行为。

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 严谨·求實·文明·公正

三、大模型数据安全保护的刑法归责理论

（二）研发者与部署者的注意义务

1.研发者应履行防控大模型安全风险的注意义务。

2.大模型算法应用后部署者应履行相应的注意义务。

3.注意义务的设定应当科学合理。

法治信仰 中国立场 国际视野 平民情怀

三、大模型数据安全保护的刑法归责理论

（三）对使用者刑事责任的区分认定

1.用户端存在“恶意”的利用和攻击行为。

2.对通常使用者（消费者）责任认定的审慎。

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 嚴謹·求實·文明·公正

四、大模型安全视域下网络犯罪的司法认定

（一）对单位犯罪的认定

各国相关产品责任与单位犯罪的立法存在不同。

1.算法研发者和部署者往往是单位（企业），涉及单位犯罪。

2.对直接责任人员的刑事归责需要慎重，决策者居于重要地位。

法治信仰 中国立场 国际视野 平民情怀

四、大模型安全视域下网络犯罪的司法认定

（二）以大模型算法为犯罪对象情形的认定

1.与传统网络犯罪相比较，侵害大模型算法安全和数据安全行为的危害性不可同日而语。

2.“提示词注入”或“模型越狱”为代表的绕过型攻击行为的认定。

3.“数据污染”或“数据投毒”为代表的数据操纵行为的认定。

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 嚴謹·求實·文明·公正

四、大模型安全视域下网络犯罪的司法认定

（三）以大模型为犯罪工具情形的认定

1.由于大模型的特质，极大地提升了法益侵害的程度，不能简单将大模型与其他犯罪工具等同视之。

2.人机协同情形下对传统的共同犯罪理论造成新的挑战。

法治信仰 中国立场 国际视野 平民情怀

大模型算法数据安全保护的刑法应对

西北政法大学　陈京春

2025年11月

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 严谨·求實·文明·公正

대규모 모델(大模型)알고리즘의 데이터 안전 보호에 대한 형법적 대응

서북정법대학(西北政法大学) Chen Jingchun(陈京春)

2025년 11월

法治信仰 中国立场 国际视野 平民情怀

一、형사법에 대한 대규모 모델 데이터 안전위험의 도전

（一）대규모성과 법익의 확정

1. 대규모 모델 알고리즘이 개인과 사회에 미치는 영향은 대규모적이다.
2. 대규모 모델 알고리즘 관련 위법행위는 사회적/공동체적 법익(集体法益)에 현실적인 침해를 가져온다.
3. 사회적/공동체적 법익에 대한 강조는 법익이 희박화(稀薄化)될 위험에 대한 경계를 요구한다.

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 嚴謹·求實·文明·公正

一、 형사법에 대한 대규모 모델 데이터 안전위험의 도전

（二）전(全) 과정성과 위험 인식

1. 대규모 모델 알고리즘의 연구 및 개발은 인위적인 안전위험을 가져온다.
2. 대규모 모델 알고리즘의 배치(部署)는 사회적 안전위험을 현실화한다.
3. 대규모 모델 알고리즘 서비스의 이용 행위 또한 법익 침해 가능성이 있다.

法治信仰 中国立场 国际视野 平民情怀

一、형사법에 대한 대규모 모델 데이터 안전위험의 도전

（三）불확정성(不确定性)과 규제 전략

1. 대규모 모델 알고리즘의 안전위험은 연구 및 개발 단계에서 충분히 인지되기 어렵다.
2. 대규모 모델 알고리즘의 안전위험은 배치 단계에서 지속적으로 드러나고 발전/변화한다.
3. 대규모 모델 알고리즘의 사용자 측 행위는 안전위험의 불확정성을 가중시킨다.

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 严谨·求实·文明·公正

一、형사법에 대한 대규모 모델 데이터 안전위험의 도전

（四）불가해성(不可解释性)과 인과관계

1. 대규모 모델 알고리즘은 일정 수준의 불가해성을 보인다.
2. 대규모 모델 알고리즘의 불가해성은 인과관계의 인정을 어렵게 한다.
3. 대규모 모델 알고리즘의 해석 가능성은 신뢰할 만한 인공지능을 위해 시급히 해결해야 할 과제이다.

法治信仰 中国立场 国际视野 平民情怀

一、형사법에 대한 대규모 모델 데이터 안전위험의 도전

（五）인간-기계 간 상호작용성과 형사귀책(刑事归责)

1. 대규모 모델 알고리즘 제품 및 서비스는 인간-기계 간 상호작용성이라는 특징을 뚜렷하게 보인다.
2. 신뢰의 원칙에 기반하여, 선의의 사용자에 대해서는 통상적으로 형사책임을 물어서는 안 된다.
3. "악의적(恶意)" 사용행위에 대하여, 어떤 경우에 책임을 물을 것인지에 대해서는 구체적인 검토가 필요하다.

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 嚴謹·求實·文明·公正

二、대규모 모델 데이터 안전 거버넌스에서 형법의 위치

（一）기술 거버넌스가 주도적 역할을 수행한다.

1. 대규모 모델 알고리즘에 대해서는 민첩한 거버넌스가 필요하며 신속하게 기술적으로 대응해야 한다.
2. 기술 거버넌스는 대규모 모델 데이터 안전 거버넌스에서 주도적 역할을 수행한다.
3. 법률규범의 내용은 기술(행위) 규범의 측면에서 구체화되어야 한다.

法治信仰 中国立场 国际视野 平民情怀

二、대규모 모델 데이터 안전 거버넌스에서 형법의 위치

（二）인공지능법은 법률체계에서 핵심적 위치에 있다.

1. 인공지능법은 행정법의 핵심이다.
2. 인공지능법은 알고리즘(데이터 처리)에 대한 안전법이기도 하다.
3. 인공지능법은 형법 개입의 경계와 척도를 탐색하기 위한 실험적 준거를 제공한다.

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 嚴謹·求實·文明·公正

二、대규모 모델 데이터 안전 거버넌스에서 형법의 위치

（三）형법적 개입과 대응은 최후수단성 원칙을 견지해야 한다.

1. 전통적인 사이버범죄의 입법 모델과 사법 활용 논리 사이에는 간극이 있다.
2. 형법적 개입과 그 형사귀책(刑事归责)에 대해서는 매우 신중한 태도를 견지해야 한다.
3. 형사입법 및 형사사법 모델의 전환이 필요한지 여부에 대해서는 심층적 검토가 필요하다.

法治信仰 中国立场 国际视野 平民情怀

三、대규모 모델 데이터 안전 보호에 관한 형법상 귀책(归责)이론

（一）다원적 주체 간의 합리적 위험 배분

1. 법규를 준수하는 연구/개발자(合规的研发者)는 허용된 위험을 창출한 것이다.
2. 배치자는 대규모 모델 알고리즘 위험의 방지/통제의 주체가 된다.
3. 보다 큰 위험은 불법적 목적에 기반한 연구, 개발 및 이용행위에서 발생한다.

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 嚴謹·求實·文明·公正

三、대규모 모델 데이터 안전 보호에 관한 형법상 귀책(归责)이론

（二）연구/개발자와 배치자의 주의의무

1. 연구/개발자는 대규모 모델 안전위험을 방지/통제하기 위한 주의의무를 이행해야 한다.
2. 대규모 모델 알고리즘의 운용 이후 배치자는 그에 상응하는 주의의무를 이행해야 한다.
3. 주의의무의 설정은 과학적이고 합리적이어야 한다.

法治信仰 中国立场 国际视野 平民情怀

三、대규모 모델 데이터 안전 보호에 관한 형법상 귀책(归责)이론

（三）사용자의 형사책임에 대한 구분적 인정

1. 사용자 측에서는 "악의적(恶意)" 이용행위와 공격행위가 존재한다.
2. 일반적인 사용자(소비자)에 대한 책임 인정은 신중해야 한다.

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 嚴謹·求實·文明·公正

四、대규모 모델 안전 관점에서의 사이버범죄에 대한 사법적 판단

（一）법인범죄(单位犯罪)의 인정에 관하여

각국의 관련 제조물책임 및 법인범죄 입법에는 차이점이 존재한다

1. 알고리즘의 연구/개발자와 배치자는 대체로 법인(기업)이므로 법인범죄(单位犯罪)와 관련이 있다.

2. 직접 책임자에 대한 형사귀책(刑事归责)은 신중하게 이루어져야 하고, 의사결정자가 중요한 지위를 차지한다.

法治信仰 中国立场 国际视野 平民情怀

四、대규모 모델 안전 관점에서의 사이버범죄에 대한 사법적 판단

（二）대규모 모델 알고리즘을 범죄의 객체로 보는 경우의 인정

1. 전통적인 사이버범죄와 비교할 때, 대규모 모델 알고리즘의 안전 및 데이터 안전을 침해하는 행위는 동일 선상에서 논할 수 없다.

2. "프롬프트 인젝션(提示词注入)" 또는 "모델탈옥(模型越狱)"과 같은 대표적인 우회형(绕过型) 공격행위의 인정.

3. "데이터 오염(数据污染)" 또는 "데이터 중독(数据投毒)"과 같은 대표적인 데이터 조작행위의 인정

西北政法大学 NORTHWEST UNIVERSITY OF POLITICS & LAW 嚴謹·求實·文明·公正

四、대규모 모델 안전 관점에서의 사이버범죄에 대한 사법적 판단

（三）대규모 모델을 범죄도구로 사용하는 경우의 인정

1. 대규모 모델의 특징으로 인해 법익침해의 정도가 대폭 증대되므로, 대규모 모델을 기타 범죄도구와 단순하게 동일시하여 처리할 수 없다.

2. 인간-기계 협동 상황 하에서, 전통적인 공범이론(共同犯罪理论)에 대해 새로운 도전이 제기되고 있다.

大模型算法数据安全保护的刑法应对

陈京春*

内容摘要：大模型算法安全风险的大规模性、全过程性、不确定性、不可解释性、人机协同性等特征，对法益保护、危险识别、应对策略、归因归责产生影响，需要构建起多元性、梯次性的安全治理体系，技术治理与行政法规范承载着重要的功能，刑法应坚守谦抑性。大模型算法的研发者在符合评估、备案和准入制度的情况下创设了法所允许的危险，在部署应用过程中危险处于持续变化状态。大模型算法的部署者承担着防控危险的重要责任。应合理设定研发者和部署者的注意义务。对大模型算法产品服务用户的刑事归责应审慎。针对以大模型算法为犯罪对象和以大模型算法为犯罪工具的情形，应对狭义的网络犯罪罪名和相关罪名的解释适用进行调整完善。对于创设相应新罪名的提议还需要冷静研究。

关键词：大模型算法；数据安全；算法安全；数据泄露；数据投毒

大模型算法数据安全风险贯穿于数据的全生命周期，包括数据采集、存储、传输、处理、应用等各个环节。大模型算法数据安全风险来自内部和外部两个方面。大模型算法内部安全治理失控可能导致数据泄露（Data Leakage）、数据污染（Data Contamination）等情形；来自外部危害大模型算法安全的攻击包括对抗样本攻击（Adversarial Attacks）、模型窃取攻击（Model Stealing）、数据投毒攻击（Data Poisoning）、后门攻击（Backdoor Attacks）等技术攻击类威胁。深度伪造（Deepfake）等对大模型算法的滥用已经成为新型网络犯罪的新形态，被社会各界高度关注。在大数据时代，针对大模型算法数据安全的风险，如何构建起安全防控机制，以及刑法的定位及其功能的发挥，成为当前重大的理论课题。

一、大模型数据安全风险对刑事法的挑战

大模型算法的独特性体现了算法研发的转向，也对刑事法律的规制带来了新的挑战。随着大模型算法在各类场景中的广泛部署，其重大的安全问题和现

* 作者简介：陈京春，西北政法大学教授、法学博士、博士生导师、数字法治与数据安全研究院院长，主要从事刑法学、数字法学研究。

实危险，对刑法的介入与功能提出了新要求。

（一）大规模性与法益的确定

1.大模型算法对个人和社会的影响具有大规模性。大模型算法的大规模性体现在语料库及其处理的大模型性、应用范围的大规模性和安全风险的大规模性。大规模性决定了涉大模型算法的法益侵害性的重大及其广泛性，不仅体现在对个人法益的侵害，更体现在对集体法益的侵害。

2.涉大模型算法违法行为对集体法益的侵害具有现实性。传统的网络犯罪，更多体现为对个人法益的侵害，而涉大模型算法的法益侵害更为凸显的是对集体法益的侵害。随着大模型算法的广泛场景应用和部署，其安全风险危及到公共安全、国家安全等维度。

3.对于集体法益的强调需要谨防法益稀薄化的风险。尽管对集体法益的强调是必要的，但是，涉大模型算法的危害行为具有抽象性和认定的困难，因此，有必要强调集体法益与个人法益的联系，以防止以集体法益为由，过度地限制大模型算法的研发与部署。

（二）全过程性与危险的识别

1.大模型算法的研发带来了人为的安全风险。由于大模型算法的自主性和算法黑箱的存在，伴随着大模型算法的优化升级，其安全风险也同时产生，这是科技进步不得不承受的代价。完全消除安全风险是不现实的，理性的目标是有效地防控安全风险。

2.大模型算法的部署使社会安全风险现实化。相比较大模型算法的研发，大模型算法的部署使得安全风险现实化为对个人和社会的危险，而且这种危险在应用推广和人机交互的过程中存在放大、变异的可能性，因此，在大模型部署应用的过程中，持续性地监测危险并敏捷地做出反应是非常重要的任务。

3.大模型算法服务的使用行为亦存在法益侵害的可能。大模型算法服务的使用端必然存在着违法使用的可能。大模型算法的机能及其原理导致其易于被非法利用或攻击，而且对于使用端的不法行为，只能进行动态的完善应对机制，猫与鼠的游戏持续进行。

（三）不确定性与规制的策略

1.大模型算法的安全风险未并在研发阶段就能够被充分认知。由于算法一

定程序的自主性，对大模型算法的安全风险进行评估，并基于结果采取备案与准入机制是必要的，规范违反说在此领域的应用越来越有力。

2.大模型算法的安全风险在部署阶段不断显现和发展变化。由于部署应用过程中可能出现投入市场之初无法完全预见的安全问题，因此部署者负有监测的义务，并需要保持相应的防控能力，在危险已经迫近或已经存在产生实害可能的情况下，应当执行算法停止程序，对相关产品进行召回处理。

3.大模型算法使用端的行为加剧了安全风险的不确定性。对使用端安全风险需要在研发和部署过程中进行全面的预判，并从技术维度和规范维护进行防范。例如对于使用端用户的不当行为，从算法伦理和技术规制上进行预设，以防止相应不利后果的发生。

（四）不可解释性与因果关系

1.大模型算法具有一定程度上的不可解释性。可解释性和透明性是可信人工智能的要求，但是，越是先进的算法越是不具有充分的可解释性。计算机科学对可解释性的关注目的与刑法规范的目的存在差异，也存在科学与规范间的密切联系。

2.大模型算法的不可解释性给因果关系的认定带来困难。算法的可解释性影响到刑法的归因归责，但是，算法是否具有可解释性只是刑法中因果关系认定需要考虑的因素之一，并不是刑法归因判断的全部，算法不具有充分可解释性并非刑法归责的鸿沟。[1]技术上算法可解释性技术的提升，将为刑法归因归责提供了更为有力的支撑。

3.在技术层面上，大模型算法的可解释性是可信人工智能亟需解决的课题。当下，计算机科学试图解决算法可解释性难题，但是，当下更为务实的策略是强调算法的可信性。刑法的视域下，大模型算法在得到指令的情况下，通常会做出什么的回应，是否符合人类的理性，似乎更为重要。

（五）人机交互性与刑事归责

1.大模型算法产品及其服务具有人机互动的显著特征。虽然人类的行为依然是刑法规范评价的对象，但是人机交互性的大模型应用场景中，需要考虑到大模型算法的介入对行为人罪过和危害行为认定的影响。

[1] 陈京春：《算法的可解释性与刑法归责》，载《法律科学(西北政法大学学报)》2025 年第 4 期，第 63-76 页。

2.基于信赖原则，善意的使用者通常情况下不应被追究刑事责任。投入市场的大模型算法产品服务的安全性应当是值得信赖的，善意的使用者基于此种信赖而实施的利用行为，即便造成一定的损害结果，也不能进行刑事归责。

3.对于“恶意”使用行为，在何种情况下需要进行追责依然需要具体考量。并非所有不当的使用行为都如“深度伪造”那样易于识别与定性。由于大模型算法的复杂性，如何认定“恶意”使用行为常常是困难的，如果背离责任主义的基本要求而刻意发挥刑法的功能，那将得不偿失。

二、刑法在大模型数据安全治理中的定位

对于大模型算法数据安全问题，需要构建起技术、伦理、规范等多元性、梯次性的安全风险防控体系，刑法在此体系中居于何种地位，需要深入探讨和理性对策。

（一）技术治理发挥着主导作用

1.由于大模型算法安全风险的重大性、大规模性、隐蔽性和互动性，需要敏捷治理，及时做出技术回应，并完善相关应对机制。大模型算法的研发者、部署者大技术上更为熟悉其内在安全风险及其运行机理，也更有能力动态地防控应用产生的安全问题。

2.与伦理治理、规范治理相比较，技术治理在大模型数据安全治理中发挥着主导作用。伦理治理的作用发挥，需要内嵌到技术层面，实现技术应用的合伦理性。包括行政法、刑法等法律规范的治理往往具有滞后性和间接性，依赖于大模型算法的研发者和部署者予以实现规范目的。

3.需要将大模型安全保护的法律规范的内容，落实到大模型研发、部署、使用的技术（行为）规范层面。虽然部分国家和地区已经制定，或正在制订人工智能性及其配套的法规规章和安全标准，但是，更为重要的问题是将这些法律规范的精神和要求贯彻到研发、部署过程性的技术规范、行为规范中，才能够现实地发挥作用。

（二）人工智能法居于法律体系的核心

1.人工智能法是行政法的核心。在大模型算法依然处于高速发展的阶段，积极推进人工智能立法，对科技的研发和产业的发展予以制度指引是非常必要的。由于行政法的属性和功能，其更有利于及时应对与调整，以适应科技发展

所带来的新问题，并且其适用的社会成本明显低于刑法的介入。

2.人工智能法也是算法（数据处理）的安全法。人工智能法及其相配套的行政法规、部门规章、安全标准及其操作制度，构成了法律治理的基本架构。在保障大模型算法为代表的人工智能产业健康发展的同时，对安全问题的强调和规制是其核心内容。安全与发展问题及其平衡是数字时代永恒的话题，二者相辅相成。

3.人工智能法为核心的行政法的立法和适用，为探寻刑法介入的边界和尺度提供了实验性的参照。基于法秩序统一性原理，行政法与刑法的衔接与配合是必要的。作为前置法的行政法的立法和执法效果，是判断刑法合理介入的前提。

（三）刑法的介入回应需坚守谦抑原则

1.由于大模型算法及其数据安全风险的固有属性，导致传统网络犯罪的立法模式与司法适用逻辑产生了罅隙。无论是对大模型算法的技术性攻击，还是内部治理存在的安全风险，在现有刑法规范解释适用的过程中，都存在适用的困境。

2.在大模型算法与数据安全风险及其行政法规制尚不明朗的情况下，刑法的介入及其刑事归责应当保持极为审慎的态度。目前，针对大模型算法的行政法立法和执法完善工作正在积极推进过程中，行政处罚的立法和适用成为新热点。在此情况下，刑法的介入应谨慎。

3.大模型算法及其数据安全风险对于刑法理论和立法的影响何其深远，是否需要转变刑事立法和刑事司法的模式，需要深入探讨。

三、大模型数据安全保护的刑法归责理论

虽然在刑事司法实践中，涉大模型算法安全的案例依然极少，但是，从现实的需求和前瞻性研究的角度看，进行刑法归因归责理论的深入研究是十分必要的。

（一）多元主体间合理的危险分配

1.从风险创造、风险实现的进程看，研发者虽然创设了人为的风险，但是，为了科技的进步和社会的发展，通常情况下这种人为的风险是被允许的风险。在此意义上讲，大模型算法一旦产生，就存在着安全风险，只不过作为产品和服务投入市场前，对这种风险的评估及其产品服务的备案、准入机制，使此类安全风险限定在国家和社会可接受的范围之内。

2.在大模型算法产品和服务投放市场之后，算法的部署者成为算法危险防控的主体（算法的研发者与部署者也可能是同一主体）。算法的部署者在研发者的协助下，应当保持防控大模型算法确定的或不确定的风险的专业能力，并在安全危险出现（或已经明确可以预判）的情况下，防范或控制此种安全危险。

3.与正常的大模型算法研发部署的市场行为相比较，更大的危险来自基于非法目的的研发、利用行为。这些非法研发、利用行为往往脱离安全制度体系的约束，制造法所禁止的危险，应当成为刑法惩治的重点。同时，市场化的大模型产品服务的研发和部署，也应考虑到被非法利用的危险，并采取相应的非法利用行为的防控和危险的阻断。

（二）研发者与部署者的注意义务

注意义务本质上是由法律所确立的能力分配方案，其目的在于针对特定的风险场域对公民有限的注意力进行合理安排。[2]由于大模型数据安全危险的全过程性和不确定性，因此需要合理地设定多元主体的注意义务。

1.研发者基于专业知识，不仅具有识别大模型安全风险的能力，亦应承担防控大模型安全风险的义务。基于当前各国大模型算法评估、备案及其产品服务准入的相关制度安排，在规范的大模型算法研发过程中，大模型算法所承载的人为安全风险能够得到一定程度的把控。对于研发者的注意义务而言，保持其应有的算法安全与数据安全治理的技术能力，遵从相关的规范要求，是其合法合规从事大模型研发活动的前提。

2.当大模型算法进入应用阶段，部署者应履行相应的注意义务。部署者应当担承结果预见义务与结果避免义务，以防止部署应用中大模型算法可能的危险转化为实害。这种结果预见义务不仅包括已经明确的大模型应用中存在的、来自于内部或外部的安全危险，也包括在监管大模型算法运行过程中认识到的（或可以预见的）安全问题。其结果避免义务包括部署者（在研发者的辅助下）采取适当的措施，阻止现实的危险转化为实害，或防控已经产生的危险和实害进一步扩大。

3.注意义务的设定应当科学合理。注意义务的设置不仅需要考虑到大模型算法带来的危险或实害的重大性，也要考虑到算法技术应对的可能性。如果某

[2] 陈璇：《社会治理视角下刑法归责模式的现代转型》，载《中国社会科学》2025 年第 7 期，第 126 页。

种大模型算法安全问题是无法从技术上解决的，而且这一安全问题对个人法益和社会法益构成重大的威胁，那么就应当禁止此类大模型算法的应用。不应设定一般的算法技术人员（甚至是卓越的从业者）不可能履行的注意义务，并因此而追究研发者、部署者的刑事责任。

（三）对使用者刑事责任的区分认定

1.用户端可能存在“恶意”地利用大模型算法或非法“攻击”大模型算法的行为。对于将大模型算法作为“犯罪工具”的场合，运用传统的犯罪罪名予以认定，如侵犯公民个人信息罪、侵犯商业秘密罪、诈骗罪等；对于将大模型产品作为“犯罪对象”的场合，可能构成非法侵入计算机信息系统罪、非法控制计算机信息系统罪或非法获取计算机信息系统数据罪等狭义的网络犯罪等狭义的网络犯罪罪名。

2.对通常使用者（消费者）责任认定的审慎。在完全由大模型算法决策的场合，诸如家用人工智能机器人、全自动驾驶汽车等应用，作为这些算法应用产品的使用者（消费者）不应承担刑法上的注意义务。在人机协同的场合，由于使用者（消费者）在与大模型算法应用的互动中可能存在不规范行为，算法研发者和运维者应针对此种风险进行机制预设，以防止人与算法应用的互动中产生客观危险或实害。

四、大模型安全视域下网络犯罪的司法认定

在大模型算法应用的时代背景下，对于狭义的网络犯罪（以网络为侵害对象实施的犯罪行为）的认定带来了诸多新问题。同时，以大模型算法为犯罪工具的危害行为的定性也存在解释适用的难题。

（一）对单位犯罪的认定

1.在市场条件下，算法研发者和部署者不排除自然人的情形，但往往是单位，存在着单位犯罪的问题。对单位犯罪进行刑法归责的前提是存在相应的单位犯罪。由于各国对于产品责任和单位犯罪的立法不同，所以，对于产品责任的承担主体也存在不同的逻辑。由于《中华人民共和国刑法》中还没有滥用算法的罪名，因此，目前只能根据已有的罪名进行犯罪认定，如拒不履行信息网络安全管理义务罪。该罪的犯罪主体是网络服务提供者，包括自然人和单位。在成立单位犯罪的情况下，实行双罚制，对单位判处罚金，并对其直接负责的

主管人员和其他直接责任人员进行处罚。

2.对直接责任人员的刑事归责需要慎重。对于高风险的大模型算法，单位（企业）决定了系统的运行与停止。虽然大模型算法运维单位（企业）内部的工作人员对算法的运维负有特定的职业上的安全监管义务，但大模型算法运用导致安全风险的认定和大模型服务的停止仍依赖于单位的决策，故一般情况下不应对算法运维单位一线的工作人员归责。但是，当主管人员或直接责任人员严重不负责任，不履行注意义务，导致重大安全事故发生的情况下，依然有刑事归责的可能性。

（二）以大模型算法为犯罪对象情形的认定

1.与传统网络犯罪相比较，侵害大模型算法安全和数据安全行为的危害性不可同日而语。虽然刑法有非法获取计算机信息系统数据罪、破坏计算机信息系统罪等狭义网络犯罪罪名，用以规制计算机领域的非法获取数据、篡改数据等行为。但是，考虑到危害大模型数据安全行为类型的多样性，现有的立法存在着“侵入+技术手段”要求过窄、“数据”范围界定不清、“情节严重”标准未能充分考虑大模型数据海量、类型多元特点的现实问题。因此对此类行为进行犯罪认定，需要进行行为类型化考察，并审慎评估现有刑法规范的适应性。

2.“提示词注入”或“模型越狱”为代表的绕过型攻击行为的认定。行为人虽未采用传统“侵入”手段，但其通过构造对抗性指令，实质上是未经授权调用或控制了计算机信息系统的特定功能，以获取本被安全策略所限制的数据或内容。此类行为给“非法获取计算机信息系统数据罪”中“侵入”或“其他技术手段”的传统解释带来疑问。可从功能主义视角将此类规避安全机制的行为解释为“其他技术手段”，同时，对“情节严重”的认定，不能仅以获取数据的数量为标准，也要考量所获取数据的敏感性、所生成内容的危害性以及对模型安全秩序的破坏程度。

3.“数据污染”或“数据投毒”为代表的数据操纵行为的认定。涉大模型算法危害行为所侵害的法益从数据秘密性转向计算机信息系统功能性法益，即模型输出准确性与公正性。此类行为危害广泛且具长期性、隐蔽性。由于其对大模型算法产品服务的计算机信息系统产生的影响，往往没有达到“破坏计算机信息系统罪”的危害后果的程度要求，所以对此类行为的认定存在困境。应

秉持刑法谦抑原则，构建“行政前置”规制路径。对于造成模型“偏差”但未产生特定严重后果的数据操纵行为，优先由行政法调整。对于那些对大模型算法及其系统造成破坏并导致功能性丧失的情形，才有动用刑法规制的必要。

（三）以大模型为犯罪工具情形的认定

1.由于大模型的特质，极大地提升了法益侵害的危害程度，不能简单将大模型与其他犯罪工具等同视之。[3]行为人利用大模型算法，可能侵害特定人的人身、财产法益，如利用“深度伪造”进行诈骗，或利用大模型算法服务侵害他人格权。同时，行为人利用大模型算法也可能危及到不特定多数人的利益，危害公共安全等集体法益。其安全风险的大规模性导致现有罪名的入罪标准和量刑标准不相适应，存在进一步完善的必要。

2.人机协同情形下对传统的共同犯罪理论造成新的挑战。在人机协同的场合，大模型算法可能成为行为人共同体建构的媒介，出现网络聚合犯罪的形态。由于各行为主体之间既缺乏意思联络，也无共同实行行为，主要体现为对违规行为、违法侵权行为甚至中立行为、生活行为的违法性聚合，具有因果关系网络稀释化、实行行为碎片化、主观罪过模糊化等特征。[4]因此，结合大模型算法参与下的网络聚合特征，合理地运用帮助行为正犯化的立法、刑事法中的推定逻辑，在遵从责任主义立场的前提下，对共同犯罪理论进行创新是有必要的。

五、余论

大模型算法的安全问题进一步提升了安全刑法的地位。面对大模型算法的安全风险，现有的罪名体系和刑法理论是否具有适应性，是刑法理论界关注的重点。随着人工智能技术的进化，人类社会逐渐从人类中心主义向人与大模型算法共治转变，刑法理论和刑事立法也随之受到前所未有的挑战。当前，赋予以大模型算法为代表的人工智能以刑事责任主体的地位，显然为时过早，也未能取得共识，但是，结合人机协同和大模型算法为链接媒介和中心的社会特征，对刑法规范进行严格限制的扩张解释，似乎成为必要之举。刑法在面对大模型算法安全风险时不能缺位，但是，刑法的过度介入也未必理性。是否需要创设新的涉人工智能（大模型算法）的罪名，还有待观察和理性思考。

대규모 모델(大模型) 알고리즘의 데이터 안전 보호에 대한 형법적 대응

첸징춘(陈京春)*

대규모 모델 알고리즘의 안전위험이 지닌 대규모성, 전(全) 과정성, 불확실성, 불가해성, 그리고 인간-기계 협동성 등의 특성은 법익 보호, 위험 인식, 대응 전략 및 귀인(归因)과 귀책(归责)에 영향을 미친다. 이에 따라 다원적이고 단계적인 안전 거버넌스 체계를 구축할 필요가 있으며, 기술적 거버넌스와 행정법 규범이 중요한 기능을 담당하게 되며, 형법은 최후수단성(ultima ratio, 谦抑性)을 견지해야 한다. 대규모 모델 알고리즘의 개발자가 평가·등록·준입(准入) 제도를 준수하는 경우 법이 허용하는 위험을 창출하는 것이며, 이 알고리즘을 배치·운용하는 과정에서 위험은 지속적으로 변화하는 상태에 놓이게 된다. 대규모 모델 알고리즘의 배치자는 위험을 방지·통제할 중대한 책임을 부담한다. 개발자와 배치자에게 부과되는 주의의무는 합리적으로 설정되어야 한다. 대규모 모델 알고리즘 제품 및 서비스 이용자에 대한 형사책임 부과는 신중을 기해야 한다. 대규모 모델 알고리즘이 범죄의 대상이 되거나 범죄 도구로 사용되는 경우, 협의의 사이버 범죄 구성요건 및 관련 범죄 구성요건의 해석·적용을 조정·보완할 필요가 있다. 이에 상응하는 새로운 범죄 구성요건(新罪名)의 신설에 관한 제안에 대해서는 보다 신중한 검토가 요구된다.

주제어: 대규모 모델 알고리즘(大模型算法); 데이터 보안(数据安全); 알고리즘 보안(算法安全); 데이터 유출(数据泄露); 데이터 중독(数据投毒)

* 서북정법대학(西北政法大学) 교수, 법학박사, 박사과정 지도교수, 디지털법치.데이터보안연구원(数字法治与数据安全研究院) 원장. 주요 연구분야: 형법, 디지털법.

대규모 모델 알고리즘의 데이터 안전 위험은 - 데이터의 수집, 저장, 전송, 처리, 활용 등을 포괄하여 - 데이터의 생애주기 전반에 걸쳐 존재한다. 대규모 모델 알고리즘의 데이터 안전 위험은 내부와 외부 양 측면에서 모두 야기된다. 대규모 모델 알고리즘 내부의 안전 거버넌스가 통제력을 상실할 경우, 데이터 유출(数据泄露, Data Leakage), 데이터 오염(数据污染, Data Contamination) 등과 같은 상황으로 이어질 수 있고; 대규모 모델 알고리즘의 안전을 위협하는 외부의 공격으로는 적대적 공격(对抗样本攻击, Adversarial Attacks), 모델 탈취(模型窃取, Model Stealing) 공격, 데이터 중독(数据投毒, Data Poisoning) 공격, 백도어 공격(后门攻击, Backdoor Attacks) 등을 포함하여 다양한 기술 기반 공격 위협들이 존재한다. 더 나아가 딥페이크(深度伪造, Deepfake)와 같은 대규모 모델 알고리즘의 남용은 이미 새로운 사이버 범죄 유형이 되었으며, 사회 각계각층으로부터 상당한 주목을 받고 있다. 오늘날과 같은 빅데이터 시대에 대규모 모델 알고리즘의 데이터 안전 위험에 대하여, 어떤 안전 통제 메커니즘을 구축할 것인지, 그리고 형법이 어떤 위치에서 그 역할과 기능을 수행할 것인지가 현재 중요한 이론적 연구과제로 자리매김하고 있다.

一. 형사법에 대한 대규모 모델 데이터 안전 위험의 도전

대규모 모델 알고리즘의 특수성은 알고리즘 연구·개발 양상의 변화를 구체적으로 드러낼 뿐만 아니라, 형사법적 규제에 대해서도 새로운 도전을 제기한다. 대규모 모델 알고리즘이 다양한 환경/영역에 폭넓게 배치되면서, 그로 인해 중대한 안전 문제와 현실적 위험이 뒤따르게 되었고, 이는 형법의 개입 방식과 기능 수행에 대해 새로운 요구가 제기되는 결과를 가져 왔다.

(一) 대규모성과 법익의 확정

1. 개인과 사회 전반에 대하여 대규모 모델 알고리즘이 미치는 영향은 대규모성(大规模性)을 특징으로 한다. 대규모 모델 알고리즘의 대규모성은 코퍼스(语料库, corpus)와 그 처리의 대규모성, 응용범위의 대규모성 및 안전위험의 대규모성에서 구체적으로 드러난다. 이러한 대규모성은 대규모 모델 알고리즘과 관련된 법익 침해의 중대성 및 광범위성을 결정짓는바, 개인적 법익의 침해에서뿐만 아니라 사회적/공동체적 법익(集体法益)에 대한 침해에서도 구체적으로 드러난다.

2. 대규모 모델 알고리즘 관련 위법행위는 사회적/공동체적 법익 침해에 있어서 현실성을 띠게 된다. 전통적인 사이버범죄는 개인적 법익 침해에 대하여 구체화되는 경우가 많았으나, 대규모 모델 알고리즘과 관련된 법익 침해는 그보다 더 두드러지게 사회적/공동체적 법익을 침해하는 양상을 보인다. 대규모 모델 알고리즘이 다양한 영역에서 광범위하게 적용·배치됨에 따라, 그로 인한 안전 위험은 공공안전, 국가안보 등 여러 측면에서 위험요소가 되고 있다.

3. 사회적/공동체적 법익에 대한 강조는 필수적이며, 법익이 희박화(稀薄化)될 위험에 대해서도 경계해야 한다. 사회적/공동체적 법익에 대한 강조가 필요함에도 불구하고, 대규모 모델

알고리즘과 관련된 위해행위는 추상성을 띠며 확정하기가 곤란하다. 따라서 사회적/공동체적 법익과 개인적 법익 사이의 관계를 강조할 필요가 있으며, 이를 통해 사회적/공동체적 법익(보호)를 이유로 대규모 모델 알고리즘의 연구, 개발 및 배치가 과도하게 제한되는 결과를 방지해야 한다.

(二) 전(全) 과정성과 위험 인식

1. 대규모 모델 알고리즘의 연구 및 개발은 인위적 안전 위험을 가져온다. 대규모 모델 알고리즘의 자주성과 알고리즘의 블랙 박스로 인해, 대규모 모델 알고리즘이 최적화되고 업그레이드됨에 따라서 그 안전위험 또한 동시에 발생하게 된다. 이는 과학기술이 발전한 데 따라 감수해야 할 대가라 할 것이다. 이러한 안전위험을 완전히 제거하는 것은 현실적으로 불가능하며, 합리적인 목표는 안전위험을 효과적으로 예방하고 관리/통제하는 데 있다.

2. 대규모 모델 알고리즘의 배치는 사회적 안전위험을 현실화한다. 대규모 모델 알고리즘의 연구 및 개발과 비교할 때, 대규모 모델 알고리즘의 배치는 개인 및 사회적 위험에 대하여 안전위험이 현실화되는 결과를 가져온다. 나아가 이러한 종류의 위험은 적용 확대, 그리고 인간-기계 간 상호작용 과정 중에 증폭되거나 변형될 가능성이 있다. 따라서, 대규모 모델 알고리즘의 배치 및 응용 과정 전반에 걸쳐, 위험을 지속적으로 모니터링하는 것과 더불어 민첩하게 대응하는 것이 매우 중요한 과제가 된다.

3. 대규모 모델 알고리즘 서비스의 이용 행위 또한 법익 침해 가능성을 내포한다. 대규모 모델 알고리즘 서비스의 이용에는 필연적으로 위법하게 사용될 가능성이 존재한다. 대규모 모델 알고리즘의 기능과 원리는 그 자체로 불법적 이용이나 공격에 취약한 구조를 형성하며, 사용자 측에서 발생할 수 있는 불법행위에 대해서는 오직 대응 메커니즘의 능동적인 개선/보완만이 가능할 뿐이다. (마치 끊임없이 반복되는) 고양이와 쥐의 유희처럼 말이다.

(三) 불확정성과 규제 전략

1. 대규모 모델 알고리즘의 안전위험은 연구/개발 단계에서 충분히 인지되기 어렵다. 알고리즘은 일정 수준의 자주성을 갖추고 있기 때문에, 대규모 모델 알고리즘의 안전위험에 대하여 평가하고, 그 결과에 기반하여 등록 및 준입(准入) 메커니즘을 채택하는 것이 필수적이다. 규범위반설(规范违反说)은 이와 같은 영역에 적용하는 데 점점 더 유력해지고 있다.

2. 대규모 모델 알고리즘의 안전위험은 배치 단계에서 지속적으로 드러나고 발전/변화한다. 대규모 모델 알고리즘이 실제로 배치/응용되는 과정에서 시장에 처음 출시될 당시에는 완전히 예견하기 어려웠던 안전 문제가 발생할 수 있다. 따라서 배치자는 이를 지속적으로 모니터링할 의무를 부담하며, 이에 상응하는 적절한 예방·통제 능력을 유지해야 한다. 위험이 이미 임박해 있거나 실제로 피해가 발생할 가능성이 있는 경우, 알고리즘 중지 절차를 실행하고 해당 제품의 회수(리콜) 등 필요한 조치를 취해야 한다.

3. 대규모 모델 알고리즘의 사용자 측 행위는 안전위험의 불확정성을 가중시킨다. 사용자

측에서 발생하는 위험에 대해서는 연구/개발 및 배치 과정 전반에 걸쳐 전면적인 예측/판단이 요구된다. 이에 더하여 기술적 차원과 규범적 보호 장치를 활용한 예방 조치가 마련되어야 한다. 예를 들어, 사용자 측의 부당행위에 대해서는, 알고리즘 윤리 및 기술 규제를 통해 사전에 대비함으로써 그에 상응하는 부정적 결과가 발생하는 것을 방지해야 한다.

(四) 불가해성(不可解释性)과 인과관계

1. 대규모 모델 알고리즘은 일정 수준의 불가해성을 보인다. 해석 가능성과 투명성은 신뢰할 수 있는 인공지능이 갖추어야 할 요건이라 할 것이다. 그러나 알고리즘이 고도화될수록, 충분한 해석 가능성을 갖추지 못하는 경우가 많다. 컴퓨터과학 분야에서 해석 가능성을 추구하는 목적과 형법 규범이 요구하는 목적에는 차이점이 존재하나, 과학과 규범 사이에는 밀접한 관련성 또한 존재한다.

2. 대규모 모델 알고리즘의 불가해성은 인과관계의 인정을 곤란하게 만드는 측면이 있다. 알고리즘의 해석 가능성은 형법상 귀인(归因)과 귀책(归责)에 영향을 주지만, 알고리즘의 해석 가능성 여부는 형법상 인과관계의 인정에 있어서 고려되는 요소들 중 하나일 뿐, 형법상 귀인(归因) 판단의 전부가 되는 것은 아니다. 알고리즘이 충분한 해석 가능성을 갖추지 못했다고 해서 형법상 귀책(归责)에 큰 문제가 생기는 것은 아니다.[1] 기술적 측면에서 알고리즘의 해석 가능성 기술이 제고될 경우, 장차 형법상 귀인(归因)과 귀책(归责)을 위해 보다 강력한 버팀목이 되어 줄 것이다.

3. 기술적 차원에서 대규모 모델 알고리즘의 해석 가능성 문제는 인공지능이 신뢰성을 갖추기 위해 시급히 해결해야 할 과제이다. 현재 컴퓨터과학 분야에서는 알고리즘의 해석 가능성 문제를 해결하기 위한 다양한 시도가 이루어지고 있으나, 현 단계에서 보다 실질적인 전략은 알고리즘의 신뢰성을 강조하는 것이다. 형법의 관점에서 보면, 대규모 모델 알고리즘이 특정 명령을 받았을 때 통상적으로 어떠한 방식으로 반응하는지, 그리고 그 반응이 인간의 합리적 이성에 부합하는지 여부가 더 중요하게 여겨진다.

(五) 인간-기계 간 상호작용성과 형사귀책(归责)

1. 대규모 모델 알고리즘 제품 및 서비스는 현저하게 인간-기계 간 상호작용적 특성을 보인다. 비록 인간의 행위는 여전히 형법 규범의 평가 대상이 되지만, 인간-기계 간 상호작용을 특징으로 하는 대규모 모델 알고리즘의 응용 환경에서는 알고리즘의 개입이 행위자의 죄과(罪过)와 위해행위의 인정에 미치는 영향을 고려할 필요가 있다.

2. 신뢰의 원칙에 비추어 볼 때, 선의의 사용자에 대해서는 통상적으로 형사책임을 물어서는 안 된다. 시장에 출시된 대규모 모델 알고리즘 제품과 서비스의 안전성은 신뢰할 만한 것이어야 하는바, 선의의 사용자가 이러한 신뢰에 근거하여 (해당 제품 및 서비스를) 이용한 경우, 비록 일정한 손해의 결과가 발생하게 하였더라도 형사책임을 부과할 수는 없다.

---

1) 陈京春,《算法的可解释性与刑法归责》, 载《法律科学(西北政法大学学报)》2025年 第4期 , 第63-76页。

3. “악의(恶意)적” 사용행위에 대하여 어떤 경우에 형사책임을 물을 것인지에 대해서는 여전히 구체적으로 고려할 필요가 있다. 모든 부당 사용행위가 “딥페이크”와 같이 쉽게 식별되고 그 성격이 규정될 수 있는 것은 아니다. 대규모 모델 알고리즘의 복잡성으로 인해 무엇을 “악의적” 사용행위로 인정할 것인지 판단하는 일은 늘상 쉽지 않다. 만약 책임주의의 기본 요구에서 벗어나 형법의 기능을 억지로 행사한다면, 얻는 것보다 잃는 것이 더 클 것이다 .

二. 대규모 모델 데이터 안전 거버넌스에서 형법의 위치

대규모 모델 알고리즘의 데이터 안전 문제와 관련하여 기술·윤리·규범 등 여러 방면에서 다원적이고 단계적인 안전 위험 방지/통제 체계를 구축할 필요가 있다. 이러한 체계 내에서 형법이 어떤 위치를 차지하느냐에 대해서는 보다 심도 있는 논의와 합리적인 대책이 요구된다.

(一) 기술 거버넌스가 주도적 역할을 수행한다

1. 대규모 모델 알고리즘의 안전위험이 지니는 중대성, 대규모성, 은폐성 및 상호작용성 등으로 인해, 민첩한 거버넌스를 필요로 하며, 이에 따라 신속한 기술적 대응과 관련 대응 메커니즘의 개선/보완이 요구된다. 대규모 모델 알고리즘의 연구/개발자와 배치자는 해당 기술의 내재적 안전 위험과 작동 원리에 대해 기술적으로 보다 잘 이해하고 있으며, 응용 과정에서 발생하는 안전 문제를 능동적으로 예방/통제할 수 있는 능력 역시 상대적으로 우월하다.

2. 윤리적 거버넌스 및 규범적 거버넌스와 비교할 때, 기술 거버넌스는 대규모 모델 데이터 안전 거버넌스에서 주도적 역할을 수행한다. 윤리적 거버넌스가 효과를 발휘하기 위해서는 기술적 층위에 내재화되어, 기술 활용이 윤리성에 부합하도록 구현되어야 한다. 행정법·형법 등을 포함한 법률규범에 기반한 거버넌스는 일반적으로 시차성(滞后性)과 간접성을 가지며, 규범 목적의 실현 또한 대규모 모델 알고리즘의 연구/개발자와 배치자에게 의존하고 있다.

3. 대규모 모델 안전 보호에 관한 법률 규범의 내용은 대규모 모델의 연구/개발·배치·이용 과정에서 기술 (행위) 규범으로 구체화될 필요가 있다. 일부 국가와 지역에서는 이미 인공지능 관련 법규와 안전 표준을 제정하였거나 제정 중에 있으나, 보다 중요한 문제는 이러한 법률 규범의 정신과 요구를 연구/개발 및 배치의 전 과정에 걸쳐 기술 규범과 행위 규범으로 구현하고 이를 실질적으로 작동하게 만드는 데 있다.

(二) 인공지능법은 법체계의 핵심을 차지한다

1. 인공지능법은 행정법의 핵심이다. 대규모 모델 알고리즘이 여전히 고속 발전 단계에 있는 현시점에서, 인공지능 관련 입법을 적극적으로 추진하여 과학기술의 연구/개발과 산업 발전에 제도적 방향성을 부여하는 것은 필수적이다. 행정법이 지니는 속성과 기능에 비추어 볼 때, 행정법은 과학기술 발전이 초래하는 새로운 문제들에 대해 보다 신속하게 대응하고 조정할 수 있다는 점에서 특히 유리하며, 그 적용에 수반되는 사회적 비용 또한 형법의 개입에

비해 현저히 낮다.

2. 인공지능법은 알고리즘(데이터 처리)에 대한 안전법이기도 하다. 인공지능법과 그에 수반되는 행정법규, 부처 규정, 안전 표준 및 운영 제도는 인공지능 분야의 법률 거버넌스를 구성하는 기본적 구조를 이룬다. 이는 대규모 모델 알고리즘으로 대표되는 인공지능 산업의 건전한 발전을 보장하는 동시에, 그 과정에서 발생하는 안전 문제에 대한 강조와 규제를 핵심 내용으로 한다. 안전과 발전, 그리고 그 균형은 디지털 시대의 영원한 화두이며, (안전과 발전) 두 요소는 상호 보완적인 관계에 있다.

3. 인공지능법을 핵심으로 하는 행정법의 입법과 적용은, 형법 개입의 경계와 척도를 탐색하기 위한 실험적 준거를 제공한다. 법질서의 통일성 원리에 비추어 볼 때, 행정법과 형법의 긴밀한 연계와 협동은 필수적이다. 전단계 법률(前置法)로서 행정법의 입법 및 집행 효과는 형법의 합리적 개입 여부를 판단하는 전제가 된다.

(三) 형법적 개입과 대응은 형법의 최후수단성(謙抑性) 원칙을 견지해야 한다

1. 대규모 모델 알고리즘 및 그 데이터 안전 위험이 지니는 고유한 속성으로 인해, 전통적인 사이버범죄의 입법 모델과 사법적 적용 논리 사이에는 간극이 발생하고 있다. 대규모 모델 알고리즘을 대상으로 하는 기술적 공격이든, 내부 거버넌스 상의 안전 위험이든 간에 모두 현행 형법 규범을 해석·적용하는 과정에서 적용에 있어 어려운 점이 존재한다.

2. 대규모 모델 알고리즘 및 데이터 안전 위험, 그리고 이에 대한 행정법적 규제가 아직 명확히 정립되지 않은 상황에서, 형법적 개입 및 형사귀책(刑事归责)에 대해서는 매우 신중한 태도를 견지해야 한다. 현재 대규모 모델 알고리즘에 대한 행정법적 입법 및 집행의 정비가 적극적으로 추진되고 있는바, 행정처벌의 입법과 적용이 새로운 쟁점으로 부상하고 있다. 이러한 상황에서 형법의 개입은 보다 신중하게 다뤄질 필요가 있다

3. 대규모 모델 알고리즘 및 그 데이터 안전 위험이 형법 이론과 입법에 미치는 영향은 지대하며, 형사 입법 및 형사 사법의 모델을 전환할 필요가 있는지에 대해서는 심층적 검토가 필요하다.

三. 대규모 모델 데이터 안전 보호에 관한 형법상 귀책(归责) 이론

비록 형사사법 실무에서 대규모 모델 알고리즘의 안전성과 관련된 사례는 여전히 극히 드물지만, 현실적 수요와 전향적(前瞻性) 연구의 관점에서 볼 때, 형법상 귀인(归因)과 귀책(归责) 이론에 대한 심층적 논의가 매우 필요하다.

(一) 다원적 주체 간의 합리적 위험 배분

1. 위험의 창출과 실현이라는 과정에서 볼 때, 비록 연구개발자가 인위적 위험을 창출한다

하더라도, 과학기술의 진보와 사회적 발전을 위해 이러한 인위적 위험은 통상적으로 '허용된 위험(允许的风险)'에 해당한다. 이러한 의미에서 대규모 모델 알고리즘은 일단 생성되는 순간부터 일정한 안전 위험을 내재적으로 포함하고 있다. 다만 해당 제품이나 서비스가 시장에 투입되기 이전에 이루어지는 위험 평가, 제품·서비스의 등록 및 준입 메커니즘을 통해, 이러한 안전위험은 국가와 사회가 용인할 수 있는 범위 내로 한정되는 것이다.

2. 대규모 모델 알고리즘 제품과 서비스가 시장에 투입된 이후에는, 알고리즘의 배치자가 알고리즘 위험의 방지/통제의 주체가 된다(알고리즘의 연구개발자와 배치자가 동일한 주체일 수도 있다). 알고리즘 배치자는 개발자의 협력을 바탕으로, 대규모 모델 알고리즘이 초래할 수 있는 확정적 위험 또는 불확정적 위험을 방지/통제할 수 있는 전문적 역량을 지속적으로 유지해야 한다. 또한 안전위험이 현실화되었거나 (이미 명확하게 예견 가능한) 경우에는 해당 안전위험을 방지/통제하기 위한 조치를 취해야 한다.

3. 정상적인 대규모 모델 알고리즘의 개발·배치에 기초한 시장 행위와 비교할 때, 더 큰 위험은 불법적 목적에 기반한 연구/개발·이용 행위에서 발생한다. 이러한 불법적 연구/개발·이용 행위는 대체로 안전 규제 체계의 통제를 벗어나 법이 금지하는 위험을 창출하므로 형법적 처벌의 중점 대상이 되어야 한다. 아울러, 시장을 통해 제공되는 대규모 모델 제품·서비스의 연구/개발 및 배치 과정에서도 불법적으로 이용될 가능성에 대한 고려가 필수적이며, 이에 대응하기 위한 불법 이용행위의 방지/통제 및 위험 차단 조치를 취해야 한다.

(二) 개발자와 배치자의 주의의무

주의의무는 본질적으로 법이 설정한 능력 배분의 구조적 설계이며, 그 목적은 특정한 위험 영역에서 시민이 보유한 제한된 주의력과 대응 능력을 합리적으로 배분하는 데 있다.[2] 대규모 모델 데이터 안전 위험의 전(全) 과정성과 불확정성으로 인해 다양한 주체들에게 부과되는 주의의무를 합리적으로 설정하는 것이 필요하다.

1. 연구개발자는 전문적 지식을 바탕으로 대규모 모델의 안전위험을 식별할 능력을 갖추고 있을 뿐 아니라, 대규모 모델 안전위험을 방지/통제할 의무 또한 부담한다. 현행 각국의 대규모 모델 알고리즘 평가, 등록, 그리고 제품·서비스의 준입(准入) 등 관련 제도에 비추어 볼 때, 정상적이고 규범화된 대규모 모델 알고리즘 연구개발 과정에서는 알고리즘에 내재된 인위적 안전 위험이 일정 범위 내에서 통제 가능하다. 연구개발자의 주의의무란, 곧 알고리즘 안전 및 데이터 안전 거버넌스를 위한 기술적 역량을 유지하고, 관련 규범과 절차적 요구에 충실히 따르는 것을 의미하며, 이는 법률과 규범에 합치하는 대규모 모델 연구·개발 활동의 전제가 된다.

2. 대규모 모델 알고리즘이 응용 단계에 진입하면, 배치자는 이에 상응하는 주의의무를 이행해야 한다. 배치자는 결과 예견 의무와 결과 회피 의무를 부담하며, 이를 통해 배치·응용 과정에서 대규모 모델 알고리즘이 초래할 수 있는 잠재적 위험이 현실적 피해로 전환되는 것

2) 陈璇：《社会治理视角下刑法归责模式的现代转型》，载《中国社会科学》2025年第7期，第126页。

을 방지해야 한다. 이러한 결과 예견 의무는, 이미 명확하게 드러난 대규모 모델의 응용 과정에 존재하는 내부 또는 외부의 안전 위험뿐만 아니라, 대규모 모델 알고리즘의 운용을 감독하는 과정에서 인식된 (또는 예견 가능한) 안전 문제까지 포함한다. 결과 회피 의무는 배치자가 (개발자의 지원 하에) 현실화된 위험이 실질적 피해로 이어지지 않도록 저지하거나 이미 발생한 위험 및 실질적 피해가 더 확대되지 않도록 방지/통제하는 조치를 취하는 것을 의미한다.

3. 주의의무의 설정은 과학적이고 합리적이어야 한다. 주의의무를 설정할 때에는 대규모 모델 알고리즘이 초래할 수 있는 위험 또는 실질적 피해의 중대성뿐 아니라, 해당 위험에 대한 알고리즘의 기술적 대응 가능성 역시 함께 고려해야 한다. 만약 특정한 대규모 모델 알고리즘의 안전 문제가 기술적으로 해결 불가능하고, 그 안전 문제가 개인적 법익 및 사회적 법익에 중대한 위협을 가하는 경우에는, 이러한 대규모 모델 알고리즘 응용을 금지하는 것이 타당하다. 또한 일반적인 알고리즘 기술자(심지어 우수한 종사자조차)가 이행할 수 없는 수준의 주의의무를 설정하여서는 안 되며, (그 불이행을 이유로) 연구개발자나 배치자에게 형사책임을 물어서도 안 된다.

(三) 사용자의 형사책임에 대한 구분적 인정

1. 사용자 측에서는 대규모 모델 알고리즘을 “악의적으로” 이용하거나, 이를 불법적으로 “공격”하는 행위가 존재할 수 있다. 대규모 모델 알고리즘을 “범죄 도구”로 사용하는 경우에는, 개인정보 침해죄(侵犯公民个人信息罪), 영업비밀 침해죄(侵犯商业祕密罪), 사기죄(诈骗罪) 등과 같은 전통적 범죄 구성요건을 적용하여 판단할 수 있다. 반면, 대규모 모델 제품을 “범죄의 대상(객체)”으로 하는 경우에는, 컴퓨터정보시스템 불법침입죄(非法侵入计算机信息系统罪), 컴퓨터정보시스템 불법통제죄(非法控制计算机信息系统罪), 컴퓨터정보시스템 데이터 불법획득죄(非法获取计算机信息系统数据罪) 등 협의의 사이버범죄에 해당할 가능성이 있다.

2. 일반 사용자(소비자)에 대한 책임 인정은 신중해야 한다. 대규모 모델 알고리즘이 전적으로 결정을 내리는 경우, 예컨대 가정용 인공지능 로봇이나 완전 자율주행 차량 등의 응용과 같은 경우에 해당 알고리즘 기반 제품의 사용자(소비자)에게 형법상의 주의의무를 부과해서는 안 된다. 반면, 인간–기계가 협업하는 경우에는 사용자(소비자)가 대규모 모델 알고리즘 응용과 상호작용하는 과정에서 비규범적 행위가 발생할 가능성이 존재한다. 이때 알고리즘 개발자 및 운영·유지관리자는 이러한 위험에 대비한 메커니즘을 미리 마련하여, 인간–알고리즘 상호작용 과정에서 객관적 위험 또는 실질적 피해가 발생하지 않도록 방지하여야 한다.

四. 대규모 모델 안전 관점에서의 사이버범죄에 대한 사법적 판단

대규모 모델 알고리즘이 광범위하게 활용되는 시대적 맥락에서, 협의의 사이버범죄(인터넷을 침해대상으로 하여 이루어지는 범죄행위)에 대한 판단에는 새로운 문제가 다수 발생하고 있다. 동시에, 대규모 모델 알고리즘을 범죄 도구로 활용한 위해행위의 법적 성격을 규정하는

데에도 해석·적용상의 어려움이 존재한다.

(一) 법인범죄(单位犯罪)의 인정에 관하여

1. 시장 조건하에서 알고리즘의 연구개발자와 배치자가 자연인인 경우를 배제할 수 없으나, 대부분은 법인(单位)에 해당하므로 법인범죄와 관련된 문제가 발생한다. 법인범죄에 대하여 형법상 귀책(归责)을 위해서는, 이에 상응하는 법인범죄 구성요건이 존재해야 한다는 점이 전제된다. 국가별로 제조물책임과 법인범죄에 관한 입법이 상이하기 때문에, 제조물책임의 귀속 주체에 대한 판단 논리 역시 서로 다르게 전개된다. 중국 형법에는 아직 알고리즘 남용(滥用算法的罪)와 같은 범죄구성요건이 마련되어 있지 않으므로, 현재로서는 기존의 범죄구성요건에 근거하여 범죄 성립 여부를 판단할 수밖에 없는데, 대표적으로 정보통신망 보안관리 의무 거부·불이행죄(拒不履行信息网络安全管理义务罪)가 이에 해당한다. 이 죄의 범죄 주체는 자연인과 법인을 포함한 인터넷 서비스 제공자다. 법인범죄가 성립하는 경우, 양벌제(双罚制)가 적용되어, 법인에 대해서는 벌금이 부과되고, 직접 책임을 부담하는 관리자(主管人员)와 그밖에 직접 책임이 있는 종업원에 대해서도 처벌이 병과된다.

2. 직접 책임자에 대한 형사귀책(刑事归责)은 신중하게 이루어져야 한다. 고위험 대규모 모델 알고리즘의 경우, 법인(기업)이 해당 시스템의 작동과 중지에 관하여 결정한다. 비록 대규모 모델 알고리즘을 운영·관리하는 법인(기업) 내부의 종사자들이 직무상 특정한 안전관리 의무를 부담하고 있다 하더라도, 대규모 모델 알고리즘 활용으로 인한 안전위험의 판단과 대규모 모델 서비스의 중단 여부는 궁극적으로 법인(기업)의 의사결정 체계에 의해 좌우된다. 따라서 일반적인 경우, 알고리즘 운영 법인(기업)의 일선 직원에게 형사책임을 물어서는 안 된다. 그러나 관리자 또는 직접 책임이 있는 사람이 중대한 직무태만을 범하고, 요구되는 주의의무를 이행하지 않아 중대한 안전사고가 발생한 경우에는, 해당 인력에 대해서도 형사책임을 부과할 가능성이 여전히 존재한다.

(二) 대규모 모델 알고리즘을 범죄 객체로 보는 경우의 인정

1. 전통적 사이버범죄와 비교할 때, 대규모 모델 알고리즘의 안전 및 데이터 안전을 침해하는 행위가 갖는 유해성은 동일 선상에서 논할 수 없다. 비록 형법에는 컴퓨터 분야에서의 불법적 데이터 취득·변조 등의 행위를 규율하기 위해 “컴퓨터정보시스템 데이터 불법취득죄”, “컴퓨터정보시스템 파괴죄” 등 협의의 사이버범죄 관련 규정이 존재한다. 그러나 대규모 모델 데이터 안전에 대한 위해행위의 다양성을 고려하면, 현행 입법에는 몇 가지 현실적 한계가 존재하는바, “침입 + 기술적 수단” 요건이 지나치게 협소하다는 점, “데이터”의 범위에 대한 정의가 불명확하다는 점, “정황이 중대한 경우(情节严重)”에 관한 판단 기준이 대규모 모델 데이터의 방대성 및 (유형의) 다양성이라는 특성을 충분히 고려하지 못하고 있다는 점 등이 그것이다. 따라서 이와 같은 행위를 범죄로 인정하기 위해서는 행위를 유형별로 고찰할 필요가 있으며, 동시에 현행 형법 규범의 활용 가능성에 대해서도 신중한 평가가 요구된다.

2. “프롬프트 인젝션(提示词注入)” 또는 “모델 탈옥(模型越狱)”으로 대표되는 우회형 공격 행위의 판단이 문제가 되고 있다. 행위자는 비록 전통적인 의미의 “침입” 수단을 사용하지 않았더라도, 대항적 지시어(对抗性指令)를 구성하여 사실상 인가되지 않은 상태에서 컴퓨터정보시스템의 특정 기능을 호출하거나 제어함으로써, 본래 보안정책에 의해 제한된 데이터 또는 콘텐츠를 획득한다. 이러한 행위는 기존의 “컴퓨터정보시스템 데이터 불법취득죄”에서 요구되는 “침입” 또는 “기타 기술적 수단”에 관한 전통적 해석에 의문을 제기한다. 기능주의적 관점에서는 보안 메커니즘을 우회하는 이와 같은 행위를 “기타 기술적 수단”으로 해석할 수 있다. 또한, “정황이 중대한 경우(情节严重)”의 판단에 있어서는 단순히 취득된 데이터의 양만을 기준으로 해서는 안 되며, 취득된 데이터의 민감성, 생성된 콘텐츠의 유해성 및 모델의 안전질서를 파괴한 정도도 고려되어야 한다.

3. “데이터 오염” 또는 “데이터 중독”으로 대표되는 데이터 조작행위의 판단이 문제가 되고 있다. 대규모 모델 알고리즘 위해행위와 관련하여 침해되는 법익은, 데이터 비밀성에서 컴퓨터정보시스템의 기능적 법익, 즉 모델 출력의 정확성과 공정성으로 전환되고 있다. 이러한 유형의 행위는 광범위한 해악, 장기성 및 은폐성을 특징으로 한다. 그러나 이로 인해 대규모 모델 알고리즘의 제품 및 서비스에 대하여 컴퓨터정보시스템에 미치는 영향은, 대체로 “컴퓨터정보시스템 파괴죄”에서 요구하는 위해결과의 정도에까지 이르지 않는 경우가 많아, 해당 행위를 범죄로 인정하는 데 어려움이 있다. 이에 형법의 최후수단성 원칙(谦抑原则)을 견지하고, “행정의 전치성(行政前置)”이라는 규제 경로를 구축할 필요가 있다. 즉, 모델에 “편차(偏差)”를 야기하였으나 특정한 중대한 결과가 발생하지 않은 데이터 조작행위는 우선적으로 행정법 규제를 통해 조정하는 것이 타당하다. 반면, 대규모 모델 알고리즘 및 그 시스템을 파괴하여 기능적 상실에 이르게 한 경우에는 형법적으로 규제할 필요가 있다.

(三) 대규모 모델을 범죄도구로 사용하는 경우의 인정

1. 대규모 모델의 경우 그 특성으로 인해 법익침해의 위해성 정도가 대폭 증대되므로, 대규모 모델을 일반적 범죄도구와 동일시하여 처리할 수는 없다.[3] 행위자가 대규모 모델 알고리즘을 이용하여 특정인의 신체적 법익 및 재산적 법익을 침해할 가능성이 있는바, “딥페이크”를 활용한 사기나 대규모 모델 알고리즘 서비스를 이용하여 타인의 인격권을 침해하는 행위 등이 여기에 해당한다. 아울러 행위자가 대규모 모델 알고리즘을 활용함으로써 불특정 다수인의 이익을 위태롭게 하거나 공공안전 등 사회적/공동체적 법익을 침해할 가능성도 존재한다. 이러한 안전위험의 대규모성으로 인해 현행 (범죄구성요건 등) 입죄기준(入罪标准) 및 양형 기준이 불일치하게 되었는바, 이에 대한 개선이 필요하다.

2. 인간–기계 협동 상황 하에서, 전통적 공범이론에 대해 새로운 도전이 제기되고 있다. 인간과 기계가 협동하는 경우, 대규모 모델 알고리즘은 행위자 간 공동체 형성의 매개로 기능할 수 있으며, 그 결과 인터넷집합범죄(网络聚合犯罪)의 출현이 가능해졌다. 각 행위 주체들 사이에 의사연락(意思联络)도 존재하지 않고, 공동의 실행행위도 없기에, 범죄는 주로 위법행위,

3) 刘宪权：《涉大模型数据犯罪刑法规制新路径》，《当代法学》2024年第6期，第13页。

침해행위, 심지어 중립적이거나 일상적인 행위까지 위법하게 집합적으로 나타나는 양상을 보인다. 이는 인과간계 네트워크의 희석화, 실행행위의 파편화, 주관적 죄책의 모호화 등의 특징을 보인다.[4] 따라서 대규모 모델 알고리즘이 참여하는 네트워크 집합의 특징을 고려하여, 방조행위의 정범화 입법 및 형사법상 추정 논리 등을 합리적으로 활용하면서, 책임주의 원칙의 준수라는 전제 하에, 기존 공범이론을 혁신적으로 재구성할 필요성이 있다.

五. 결론

대규모 모델 알고리즘의 안전 문제는 안전형법의 위상을 한층 격상시켰다. 대규모 모델 알고리즘의 안전위험에 대하여 현행 범죄 구성체계와 형법 이론이 과연 적합성을 갖추고 있는가는 형법학계가 주목하는 핵심 쟁점이다. 인공지능 기술의 진화와 함께 인간 사회는 점차 인간 중심주의에서 인간과 대규모 모델 알고리즘의 공동 거버넌스(共治)로 전환되고 있으며, 이에 따라 형법 이론과 형사 입법도 전례 없는 도전에 직면하고 있다. 현재 대규모 모델 알고리즘을 대표로 하는 인공지능에 대해서 형사책임 주체로서의 지위를 부여하는 것은 시기상조이며, 이에 대한 공감대도 아직 형성되지 않았다. 그러나 인간-기계의 협동과 대규모 모델 알고리즘이 결합하여 사회적 연계의 매개이자 중심축으로 기능하는 특성을 고려하면, 형법규범에 대한 엄격한 제한 아래 확장해석을 모색하는 것이 필요해 보인다. 대규모 모델 알고리즘의 안전위험에 직면하여 형법이 부재(缺位)할 수는 없지만, 그렇다고 형법의 과도한 개입이 반드시 합리적이라고 할 수도 없다. 인공지능(대규모 모델 알고리즘)과 관련하여 새로운 범죄 구성요건을 창설할 필요가 있는지 여부에 대해서도 신중한 관찰과 합리적 사고를 요한다.

4) 于冲：《论网络聚合犯罪的刑法规制》，载《中国法学》2025年第5期，第164页。

Discussion Paper on Session 1

Crime Prevention and Data Security in the AI Era

Ko, Myoung-su

Professor, Seoul National University Law School

Jo, Hyoung Chan

Research Fellow, Korean Institute of Criminology and Justice

토론문

[제1주제] AI 시대의 범죄예방과 데이터 보안에 대한 토론문

고명수
(서울대학교 법학전문대학원 조교수)

첨단기술과 형사법 국제세미나에 함께 할 수 있어 영광입니다. 특히 Erich Marks 독일 범죄예방대회 대표님, 陈京春 중국 서북정법대 교수님의 발표를 듣고 관련 내용을 함께 고민해 볼 수 있어 기쁘게 생각합니다. Erich Marks 대표님은 인공지능이 우리 사회에 미칠 영향을 총망라하여 정리해 주셨고, 특히 범죄 예방을 위해 인공지능을 어떻게 활용할 수 있는지, 그리고 그 과정에서 발생할 수 있는 문제를 어떻게 통제하는 것이 바람직한지에 대해 논하셨습니다. 2026년 4월에 열릴 제31회 독일 범죄예방대회가 무척이나 기대됩니다. 陈京春 교수님은 대규모 모델 알고리즘 관련하여 어떠한 위험 요소가 있는지를 분석하고, 이 위험을 통제하는 수단 중 하나인 형법의 역할에 관해 논하셨습니다. 또한. 대규모 모델 알고리즘 관련 범죄 현상, 형사 책임 귀속을 둘러싼 이론적 쟁점 및 예상되는 가벌성의 공백을 정리해 주셨습니다. 많은 것을 배우고 고민해 볼 수 있었습니다. 감사드립니다. 저는 토론자로서, 두 분의 예상 또는 주장을 보다 명확히 하고, 두 분이 던진 물음에 대해 함께 고민해 보고자 합니다.

Erich Marks 대표님은 AI 기술 활용으로 인한 나타날 수 있는 여러 문제를 (형)법을 통해 해결하기보다, 도덕적 해결, 사회 규범적 해결, 또는 자율적 해결을 우선시하고 있고, 陈京春 교수님도 다원적이고 단계적인 안전 거버넌스 체계를 구축하려면 형법보다는 행정법적 규율이 타당하다고 보시는데, 저도 전적으로 동의합니다.

범죄 예방은 이제 AI 기술 발전에 힘입어 사회·경제 정책적 차원으로까지 확장되어 범죄 발생의 구조적 요인을 줄이는 데 초점이 맞춰져 있습니다.[1] AI 기술은 특정 행위 및 행위자의 통계적 위험성을 정량화하여 범죄 예방의 효율성을 높여 줍니다. 그러나 통계

1) 이른바 3차적 예방이라고 하여, 일반예방, 특별예방을 포함하여 억압을 통한 예방(재범 방지)도 포함하는 개념으로 확장되고 있습니다. 이에 대한 상세는 Kaiser, ZRP 2000, 151 (156) 참조.

적 위험성을 이유로 이른 시점에 취해진 국가의 개입 조치는 필연적으로 불평등, 자의성, 예측 불가능성 문제를 낳고[2] 시민의 기본권을 실질적으로 제약합니다. 이러한 갈등을 최소화하기 위해 전통형법과는 구별되는 개입형법(Interventionsstrafrecht)[3]의 영역이 자리 잡으면서 이상(異常) 행위들을 매우 이른 시점에 통제하게 되지 않을까 예상해 봅니다. 다만, 통계적 위험성 평가 과정에서, 통계의 특성상 차별받는 집단이 없을 수는 없겠지만, 가능한 한 데이터의 공정성, 중립성을 제고하기 위해 노력하여야 할 것입니다.

Erich Marks 대표님은 "예방 활동에서의 기회와 책임 있는 활용"(ppt 33면)에 대해 말씀해 주셨습니다. 이 맥락에서, AI 기술을 단순히 통계적 위험성을 정량화하는 데에만 활용할 것이 아니라, 객관적·실질적인 범죄 상황을 분석·평가하기 위한 도구로 적극 활용할 필요가 있습니다. 범죄 통계의 분석 대상을 다양화하고, 분석의 정확성을 높여야 합니다. 매스컴을 장식하는 범죄 뉴스로 인해 시민들은 범죄에 대해 막연한 두려움을 갖고 있고, 이 두려움은 갈수록 증폭되고 있습니다. 시민들의 범죄에 대한 인식으로서의 지각(知覺)이 객관적·실질적인 범죄 상황과 현저히 엇갈리는 현상은 전 세계에서 확인됩니다.[4] AI 기술이 이 문제를 해결하는 데 유용하게 사용되었으면 합니다.

다음으로, AI 기술과 관련된 범죄 현상, 형사책임 귀속을 둘러싼 이론적 쟁점 및 예상되는 가벌성의 공백에 관해 이야기 나누고 싶습니다.

실제 많은 사람들이 인간이 AI 기술을 제어하지 못하게 되지 않을까 걱정하고 있습니다. 과거 AI 기술이 if-then 규칙 기반 시스템이었는데, 이제는 인간이 해석할 수 있는 규칙을 거쳐 결론에 도달하는 것이 아니기 때문, 즉 해석 가능성이 없기 때문으로 보입니다. 그러나 스탠퍼드대 앤드류 응(Andrew Ng) 교수는 이러한 걱정이 화성에 인구가 너무 많아질 것을 걱정하는 것과 다를 바 없다며 AI 기술에 공포를 느끼는 것은 불필요하다고 힘주어 말합니다. AI 기술은 매우 빠른 속도로 발전하고 있습니다. 과거 인스트럭트GPT와는 달리 챗GPT 안전모듈은 사용자 프롬프트의 오류를 파악할 수 있게 되었습니다. OpenAI는 인스트럭트GPT를 만들 때보다 데이터수집 설정을 보완하였고, 데이터 작업자들의 인종 및 성별을 다양화하여 데이터의 공정성, 중립성을 높였으며, API를

2) Kaiser, ZRP 2000, 151 (155).
3) Hassemer는 현대 사회의 여러 문제는 형법과 질서위반법 사이 또는 민법과 공법 사이에 위치하는 개입법에서 규율하는 것이 바람직하다고 보았습니다. 개입(간섭)법상 제재는 형법의 기본 원칙에 비해 완화된 기준 및 절차에 의해 부과되지만, 형벌에 비해 가혹하지는 않습니다.
4) Kaiser, ZRP 2000, 151 (153 f.); Kaiser는 이러한 지각은 인위적인 산물, 즉 대중매체 또는 이익집단에 의해 부추겨지거나 왜곡된 감정일 수 있다고 지적합니다.

도입하여 폭력적·성적 내용을 사전 분류 및 제어하고 위험한 내용이나 부적절한 요청은 거부할 수 있게 되었습니다.[5] 현재 대규모 모델 알고리즘 서비스는 주요 IT 거대기업이 개발·관리·통제하고 있고, 고품질의 데이터를 사용하고 있으며 안전 모듈을 크게 강화하고 있습니다.[6]

陈京春 교수님은 배치자로 하여금 안전 문제를 지속적으로 모니터링 하도록 의무를 부과하는 것이 중요하다고 하시면서, "위험이 이미 임박해 있거나 실제로 피해가 발생할 가능성이 있는 경우, 알고리즘 중지 절차를 실행하고 해당 제품의 회수 등 필요한 조치를 취해야 한다"(발표문 3면)고 하셨는데, 이때 위험성의 판단 주체 및 기준을 어떻게 설정하는 것이 좋을지 고견을 청합니다. 우리가 기존의 법현상과 달리 AI 기술에 결부된 위험을 특별히 고민하는 이유는 알고리즘의 블랙박스 현상 때문입니다. 실제 위험이 실현되고 그와 관련된 위험 요소가 밝혀진 경우에는 당연히 이 위험 요소를 차단·관리하여야 하는 의무를 개발자/배치자에게 부여하겠지만, 관련 위험 요소가 알려지지 않은 위험 발생 가능성 단계에서는 이러한 요구에 대한 기대가능성이 없기에 형사 책임을 지울 수 없습니다. 그럼에도 불구하고 이 단계에서도 배치자에게 형사 책임을 지울 필요가 있다고 보시는지, 그렇다면 어떤 방식을 취하여야 할지, 그리고 위험성 판단의 기준을 어디에 두어야 할지에 관해 고견을 청합니다.

교수님은 선의의 사용자에 대해서는 신뢰 원칙을 이유로 형사 책임을 지워서는 안 된다고 주장하시는데(발표문 4-5면), 일단 과실범에서의 신뢰 원칙은 해당 사회 교류에서 요구되는 주의의무에 적합하게 행위한 자는 다른 참여자 또한 자신과 같이 주의의무에 적합하게 행위할 것을 신뢰하여도 좋다는 것인바, 대규모 모델 알고리즘 제품과 서비스의 안전성을 신뢰하여도 좋다는 차원의 개념이 아닙니다. 그리고 제시하신 사용자의 선의/악의 개념에 대한 부연 설명을 부탁드립니다. 교수님은 악의적 사용행위를 부당 사용행위로 이해하시는 것 같은데, 그렇다면 악의는 대규모 모델 알고리즘이 도출한 결과가 허위 또는 왜곡된 정보임을 안 경우가 아니라, 대규모 모델 알고리즘을 통해 자신이 원하는 결과를 의도적으로 도출한 경우를 말씀하시는 것 같습니다. 악의적 사용행위의 예로 딥페이크를 제시해 주셨는데, 그 외에 어떤 행위가 있을 수 있을까요. 그에 반해 선의는 대규모 모델 알고리즘 결과를 믿고 어떠한 행위를 한 경우일 것인데, 한국 판례는 변호사 등 전문가의 자문을 신뢰하고서 한 행위에 대해서도 법률의 착오(금지착오)를 인

5) 박상길, 비전공자도 이해할 수 있는 AI 지식, 비즈니스북스, 341-342면.
6) 박상길, 비전공자도 이해할 수 있는 AI 지식, 비즈니스북스, 344-350면.

정하지 않습니다.[7] 중국도 다르지 않을 것 같은데, 그렇다면 대규모 모델 알고리즘이 제시한 결과를 신뢰하고서 한 행위가 예외 없이 면책 대상이라고 보는 것은 무리가 있다고 생각합니다.

발표문 7-8면에서는 단계를 나누어 개발자/배치자의 주의의무의 일반을 제시해 주셨습니다. 그런데 대규모 모델 알고리즘을 둘러싼 위험 요소를 완전히 파악할 수 없는 관계로 위험 관리 의무 설정은 쉽지 않습니다. 이러한 이유로 기본적으로 범죄 구성요건을 고의범이자 추상적 위험범으로 만들어야 할 것인데, 분명하게 확인되지 않은 위험 요소까지 포괄적으로 포함될 수 있게 구성요건을 만들면 명확성 원칙 위반 소지가 있을 것입니다. 결과 발생에 대한 객관적 예견가능성이 없는 경우가 대부분이어서 과실범 처벌도 쉽지 않습니다. 기술이 차차 발전함에 따라 부과 의무가 구체화 되고 허용된 위험의 범위도 조금씩 축소되겠습니다만, 그때까지는 객관적으로 예견할 수 없는 위험 영역은 - 사회적 필요성 때문에 허용된 위험으로 구성해 놓은 이상 - 형사 처벌하여서는 안 될 것입니다.

발표문 10면에서 교수님은 데이터 조작행위는 행정법을 통해 우선 규율하고, 경우에 따라 형법을 통해 규율할 필요가 있다고 하셨는데, 대규모 모델 알고리즘이 학습할 수 있는 데이터라면 그것을 조작하는 모든 행위를 국가가 규율하는 것은 - "표현의 자유"를 고려할 때 - 과잉 규율이 아닌가 하는 생각이 듭니다. 허위의 사실을 인터넷에 올려놓아 그로써 타인의 명예를 훼손하거나 타인의 업무를 방해한 경우는 형사 처벌할 수 있습니다. 이에 해당하지 않는 경우는 행정법적 규율이 필요하다는 것인데, 물론 대규모 모델 알고리즘이 허위 정보를 학습하고 이를 널리 알리게 되면 광범위하고 예측 불가능한 사회적 위험을 초래할 수 있습니다. 즉 시스템적 위험과 정보 환경의 공익성을 보호할 필요는 있겠습니다만, 한 개인이 자신의 인터넷 블로그에 허위의 사실(데이터)을 기재한 행위에 대해, 설령 대규모 모델 알고리즘이 해당 데이터를 학습하게 될 것을 알고서 하였더라도, 국가가 개입하는 것은 비례 원칙에 부합하지 않습니다. 애초에 대규모 모델 알고리즘은 한 개인이 올려놓은 부정확한 또는 허위의 정보를 걸러낼 수 있어야 합니다. 이것이 가능하여야만 대규모 모델 알고리즘은 단순 사유재가 아닌, 사회의 핵심 정보 인프라로 인정받을 수 있습니다. 그래야 대규모 모델 알고리즘의 시스템적 위험과 정보 환경의 공익성 보호를 위한 일련의 국가제재가 정당화될 수 있습니다.

---

7) 대법원 1990. 10. 16. 선고 90도1604 판결; 대법원 1995. 7. 28. 선고 95도702 판결 등.

그에 반해 허위성을 인지하고 “대규모로” 시스템을 오염시키려는 행위 방식은 규율이 필요합니다. 대규모의 조직적 행위가 전제되어야 할 것입니다.
이에 대해 교수님은 어떻게 생각하시는지, 그리고 데이터 조작행위 중 행정법적 규율 대상이 되는 행위와 그렇지 않은 행위 간 경계를 어떻게 설정하는 것이 좋을지 여쭙고 싶습니다.

같은 맥락에서 발표문 11면에서 인간-기계 협동 상황에서 전통적 공동범죄 이론의 재구성 부분도 생각해 볼 필요가 있습니다. 가벌성의 공백이 있더라도 일반 공범이론을 대규모 모델 알고리즘 관련하여서는 예외를 인정하겠다는 것은 받아들이기 어렵습니다. 교수님도 같은 이유에서, 공범이론에 따른다면 방조에 불과한 행위도 그 행위 자체를 구성요건으로 하는 입법을 하고, 간접사실에 따른 고의의 우회적 추론을 상대적으로 쉽게 하는 방법을 취할 것을 제안하신 것 같습니다. 그러나 의사연락 및 공동의 실행행위가 없음에도 공동정범 이론을 ‘혁신적으로 재구성’하면서까지 형사 책임을 지우는 것이 타당한 것인지는 의문입니다. 시스템적 위험과 정보 환경의 공익성 보호를 위해 대규모 모델 알고리즘의 시스템 그 자체를 보호하기 위한 범죄구성요건을 신설하더라도, 공동의 범행 결의가 없는 자를 공동정범으로 처벌하는 것은 정당화될 수 없습니다. 공동정범이 아닌 단독정범 형태로 해당 행위를 한 것만으로는 인과관계가 규명될 수 없고 미수범 처벌 규정을 두고 있지 않아 불가벌인 경우가 있다면, 이는 형법의 단편성에 해당하는 부분입니다. 공범 이론을 대규모 모델 알고리즘에 한정하여 재구성하는 것은 흡사 빈대 잡으려다 초가삼간 태우는 격일 수 있습니다. 교수님이 생각하시는, 예외적으로 대규모 모델 알고리즘 관련하여서는 공동정범 법리가 수정되어야 하는 경우는 구체적으로 어떤 경우인가요.

마지막으로 Erich Marks 대표님이 제시한 AI 및 예방 분야에서의 국제 협력 과제(ppt 40면)도 훗날 국제 분쟁으로까지 이어질 수 있는 매우 중요한 문제라고 생각합니다. AI 기술을 통해 타국의 정보를 수집·활용하는 과정에서 발생할 수 있는 갈등 요소 및 그에 대한 합리적인 대응 방안에 대해 대표님의 고견을 청합니다.
유럽연합은 매우 모범적인 포괄적인 AI 규제법(EU AI Act)을 제정하였습니다. 뒤이어 한국은 AI 기본법(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법)을 제정하였는데, 이 법률이 담고 있는 중복·과잉 규제를 비판하는 목소리가 거셉니다. 세계 주요 국가들이 AI 주도권을 두고 경쟁하는 상황에서, 과도한 규제보다는 규제의 최소화를 통해 기술

개발을 장려하여야 한다는 주장이 주를 이룹니다. 이러한 주장은 중국의 AI 대응을 참고한 것인데, 중국은 과도한 규제로 AI 기술 발전을 저해하는 것도 안보 위협이라고 인식하고서 포괄적 AI 규제 입법에는 신중하되, 데이터 3법을 중심으로 관련 법령 간 정합성 제고 및 입법 공백을 보완하는 방식으로 대응하고 있고,[8] 미·중 기술 패권 경쟁 속에서 글로벌 AI 거버넌스 이니셔티브를 제안하는 등 AI 패권을 잡기 위해 노력하고 있습니다.[9] AI 산업은 거대 산업으로 국운이 달렸다고 해도 과언이 아닙니다. 국제 안보 차원에서도 중요한 영역입니다. 따라서 주요국은 서로 패권을 쟁취하기 위해 치열하게 다툴 것입니다.

이에 대한 해결 방안 모색 차원에서, 유럽연합 회원국 간 규제 적용 과정에서 법적·문화적 조건이 서로 달라 발생한 문제는 없었는지, 유럽연합 국가와 거래하는 유럽연합 비회원국이 EU AI Act 규제를 수용하는 과정에서 확인된 갈등 요소는 없었는지, 그리고 그것을 해결하기 위해 어떠한 노력을 하고 있는지 궁금합니다.

이상으로 토론을 마치겠습니다. 감사합니다.

8) 이상우, 중국의 인공지능・데이터 입법 동향과 시사점, 동북아법연구 제19권 제1호, 1면 이하 참조.
9) 박강민, 장진철, 안성원, 유럽연합 인공지능법(EU AI Act)의 주요내용 및 시사점, 소프트웨어정책연구소, 2024. 8. 6., 22면.

토론문

“예방 분야에 있어서의 인공지능”에 관한 토론문

조형찬
(한국형사·법무정책연구원 부연구위원)

Marks 대표님께서는 발제문을 통해 독일범죄예방대회(DPT)에서 진행해 온 범죄예방에 있어서의 인공지능과 관련된 논의의 현황과 앞으로의 논의 방향 등을 매우 일목요연하게 정리해주셨으며, 발제문의 제목은 범죄예방을 전제로 작성되었지만, 그 내용은 형법뿐만 아니라 헌법상 기본권 주체나 사법제도에 있어서도 인공지능이 대두되는 현상에 어떻게 대응하는 것이 좋을지에 대한 향후 논의의 방향 등에 대해서도 여러 시사점을 제공했다고 생각합니다. 특히 인공지능을 단순한 기술이 아닌 ‘사회적 책임의 도구’로 인식해야 한다는 슬라이드 34면에 언급된 결론은 인공지능이 우리 사회에 미치는 영향에 어떻게 대응해야 하는지를 매우 압축적으로 잘 표현하셨다고 봅니다.

슬라이드 20면에서 소개된 바와 같이 독일 내 인공지능 관련 팟캐스트 서비스가 다수 존재한다는 것은, 인공지능이 초래할 세상의 변화에 많은 사람들의 관심이 집중되고 있음을 방증하는 것이라 생각합니다. 이미 그러하고 있지만, 모든 법영역의 법학자들이 인공지능과 법과의 관계에 더욱 많은 관심을 가져야 할 것입니다. 이하에서는 발제문과 관련된 질문과 추가 코멘트를 적어 드립니다.

슬라이드 22면에 언급된 바와 같이, 인공지능은 엄청난 기회를 제공함과 동시에 상당한 도전 과제를 부여하고 있습니다. 이는 ‘직역’의 존립과도 연결되는 지점이기도 한데, 슬라이드 24면에서 Vals Legal AI Report(VLAIR)에 따르면 인공지능이 변호사보다 더 나은 경우도 있다는 점이 언급되어 있습니다. 한국에서도 법률문제에 직면했을 때 일단 ChatGPT 등 생성형 인공지능 서비스를 통해 사전 지식을 얻고 이를 나름대로 교차검증을 한 뒤 변호사 상담을 하는 경우도 많다는 이야기가 있는데, 독일에서는 인공지능의 등장 및 발전으로 인해 법조 직역이 어떠한 위기에 직면해 있거나 어떠한 변화가 이루어지고 있는지를 여쭙니다.

슬라이드 31면에 언급된 바와 같이, 생성형 인공지능을 활용한 각종 허위 정보가 만연해 있습니다. 이는 상업용 광고, 선거 기간 중 특정 후보자와 관련된 영상, 그 외에도 다양한 형태로 생성형 인공지능이 만들어낸 정보들이 존재합니다. 이에 대해서는 '언급된 정보의 내용이 허위인지', 그리고 '그 영상에 담긴 인물이 실제로 그러한 발언을 하였는지' 등 허위의 대상이 다양하게 나타날 것입니다.

물론, 이에 대해 기본적으로 인공지능 활용 시 활용 사실에 대한 표시의무를 부여함으로써 일정 정도 제약이 있을 수 있겠지만, 인공지능을 활용했다는 사실을 알리지 않는 것과 그에 담긴 정보가 허위라는 점은 다른 차원의 문제이므로, 생성형 인공지능에 의해 형성된 허위 정보에 기존의 사기죄 등의 법리를 적극적으로 활용하여 대응할 수 있는 것인지, 아니면 새로운 입법 등이 필요한 영역인지를 여쭙니다.

슬라이드 33면에 언급된 바와 같이, 인공지능은 인간의 판단력을 대체해서는 안 되며 책임감 있게 사용되는 도구여야 한다는 점에 깊이 공감합니다. 한국에서도 매우 가벼운 대화나 예능 프로그램의 여러 장면에서도 생성형 인공지능에게 무언가 질문을 던져 그로부터 산출된 결과를 일종의 결론으로 보며 재밌게 화제를 이끌어가는 모습을 심심치 않게 볼 수 있습니다. 더 나아가, 실제로 변호사가 법정에 제출할 서면에 인공지능을 통해 생성된 판결 정보를 교차검증 없이 기재하여 제출하였다가 그 판결이 존재하지 않는 것으로 드러나 법정 제재를 받는 경우도 있는 것으로 알려져 있습니다. 이는 매우 단편적인 예이지만, 실제 어떤 문제에 부딪혔을 때 비용 절감이나 편의성 등을 이유로 인공지능의 산출 결과에 쉽게 의존하게 되면, 인간의 판단 자체를 대체하는 것이 일상화될 위험이 더욱 커질 것입니다.

인공지능 기술의 발전에 대한 규범적인 대응과 더불어, 발제문에서 여러 차례 강조하신 여러 학제 간 혹은 시민사회와의 협력을 통한 사회적 통제 방안을 모색하는 것도 중요하고, 또한 인공지능의 '적절한' 활용을 위한 시민의식의 제고도 절대적으로 필요하다고 생각합니다. 본 토론자도 향후 DPT의 논의를 지켜보며 인공지능의 발전에 어떻게 대응하는 것이 바람직한지에 대한 논의에 더욱 깊이, 그리고 구체적으로 참여할 수 있도록 노력하겠습니다. 뜻깊은 발제를 해주신 대표님께 거듭 감사드립니다. 〈끝〉

토론문

“대규모 모델 알고리즘의 데이터 안전 보호에 대한 형법적 대응”에 관한 토론문

조형찬
(한국형사·법무정책연구원 부연구위원)

첸징춘 교수님께서 본 발제문을 통해 ‘대규모 모델 알고리즘’ 관련 위법행위의 규율 방식에 대해 거시적인 부분부터 미시적인 부분까지 문제 될 수 있는 법적 쟁점들을 매우 잘 정리해주셨습니다. 침해법익의 획정, 형사책임 귀속 주체, 행정벌 전치주의의 필요성 및 형법의 최후수단성에 따른 적용 법역의 구분, 공범이론의 적용 가능성 등, 실무에도 상당한 영향을 미칠 여러 법이론적인 쟁점들을 대규모 모델 알고리즘의 특성에 따라 경우의 수를 분류하고 각 부분별로 교수님의 견해까지 제시하시는 등, 내용의 구체성이 매우 두드러져 발제문을 읽는 것만으로도 매우 많은 공부가 되었습니다.

특히 글 전반에 걸쳐 기술의 발전이 지니는 장점 및 단점을 명확히 구분하고, 인공지능 산업의 건전한 발전과 안전 등을 중심으로 하는 공공의 이익 보호가 규범적으로 어떻게 조화를 이루는 것이 바람직할지에 대한 깊은 고민이 담겨 있는데, 토론자인 저로서는 추상적으로만 생각해오던 것이 이 글에 구체적으로 담겨 있다는 점에서 그 통찰력에 깊은 감동을 느낍니다. 좋은 발제에 거듭 감사드립니다. 이하에서는 교수님의 발제문에서 궁금한 점과 추가 코멘트를 적어 드립니다.

발제문 一.의 (五) 부분에서 ‘인간-기계 간 상호작용’을 특징으로 하는 대규모 모델 알고리즘 응용 환경에서의 형사책임의 귀속과 관련하여 “신뢰의 원칙”에 비추어 ‘선의’의 사용자에 대해서는 통상적으로 형사책임 추궁을 물어서는 안 되고, ‘악의적인’ 사용에 대해서는 형사책임을 물을 수 있는 범위를 판단할 때에는 사안별 구체성을 고려해야 한다는 내용이 포함되어 있으며, 이에 대해서는 기본적으로 매우 동의합니다. 특히 ‘악의적인’ 사용에 대해서는 딥페이크 등과 같이 공공의 법익침해를 유발하는 것이 분명한 경우에 한하여 형법적 규율을 가하고, 그러하지 않은 경우에는 글 전체의 취지에 비추어 행정벌 등의 적용 가능성을 고려해야 할 수도 있겠습니다.

다만, ‘선의’와 관련하여 한 가지 질문이 있습니다. 시장에 출시된 제품과 서비스의 안전성에 대해서도 어느 정도의 신뢰가 있어야 ‘선의’라고 할 수 있는지 기준선을 정립하는 것이 쉽지 않아 보입니다. 이는 제품과 서비스 배포자의 설명의무와도 관련이 있을 것이고, 그러한 배포자의 설명의무가 온전히 이행되었더라도 서비스 이용에 따른 결과가 일정 정도 부당하더라도 선의라고 해석해야 하는 규범적인 상황이 있을 텐데, 교수님께서 생각하시는 ‘선의의 사용’의 예시로 무엇을 들 수 있을지를 여쭙니다.

발제문 二. 부분 전체와 四.의 (二) 부분에 걸쳐 형법의 최후수단성을 일관되게 강조하시면서 제재의 필요성이 있으면 우선적으로 행정벌 제도를 활용하는 “행정전치” 방식의 입법을 취해야 한다는 내용이 있으며, 기본적으로 매우 동의합니다. 법원의 재판을 통해 확정된 후에 본격적인 집행이 이루어지는 형벌과 달리, 과태료 등과 같은 행정벌은 일단 행정청이 시정조치 등을 한 후 그 미이행에 따라 부과되는 것으로서 처분을 받은 당사자는 행정소송 등의 형식으로 사후적 불복을 통해 다투게 되므로, 신속한 제재의 필요성이 있는 사안에 대해서는 행정청의 일정 정도 재량에 따라 선제적 대응이 가능하다는 점에서 형벌과 또 다른 실효(實效)적인 수단이기도 할 것입니다.

다만 최근 들어 한국에서도 마찬가지로, 대규모 모델 알고리즘 등 인공지능 관련 영역에서뿐만 아니라 사회 영역 전반에 걸쳐 행정벌의 입법이 매우 적극적으로 이루어지고 있으며, 이는 입법자의 입장에서도 벌칙의 도입에 비해 상대적으로 부담이 덜하기 때문인 것으로도 보입니다. 하지만 행정제재 조항의 범람도 또 다른 사회적 비용을 초래할 수 있는 바, 행정벌의 도입에 대해서도 구체화된 입법 기준과 사회적 합의를 전제로 이루어지는 것이 바람직하다는 생각이 들었습니다.

마지막으로, 이는 제재를 위한 구성요건의 조건과도 관련된 내용일 텐데, 발제문의 二.의 (一) 부분에서 대규모 모델 안전 보호에 관한 법률 규범의 내용을 연구/개발 – 배포 – 이용 등의 과정에서 기술 규범과 행위 규범으로 구현하고 이를 실질적으로 작동할 수 있도록 설계해야 한다는 점을 지적하셨습니다. 말씀하신 바와 같이 대규모 모델 알고리즘의 활용 방식에 따라 제재가 필요한 행위 태양은 매우 다양하고, 태양별로 많은 차이가 나타날 것이므로 구체성을 띄어야 한다는 점에 매우 동의합니다.

다만 그것이 제재 조항의 구성요건으로 규범화되는 과정에서 특정 기술에 대한 대응을 이유로 구체성이 과도해진 형태로 입법된다면, 과학기술 관련 행정법제나 지식재산법제 등에서 일부 나타나는 것과 같이 법문언상 ‘기술중립성’이 발현되지 못할 수도 있고, 또한 규정 형식의 복잡성으로 인해 수범자들이 어떠한 행위가 금지되는지를 파악하기 어렵게 될 수도 있으므로, 구체성의 정도에 대해서도 신중한 접근이 필요해보입니다.

인공지능 기술의 발전으로 예기치 못한 현상들이 빠르게 나타나고 있으며, 그러면서도 법이라는 속성상 사회적 합의를 통해 규범화 작업이 필요하다는 점에서, 대응의 시간적 괴리가 생기는 것은 필연적이라고 생각합니다. 다만, 그러한 점을 이유로 사회 현상에 드러난 문제점을 방치하는 것은 옳지 못하므로, 교수님의 발제문과 같은 구체적이고 다양한 쟁점의 논의를 적극적으로 이끌어 내어 사회적 합의의 속도를 빠르게 진척시키는 것이 우리 사회의 문제에 법이 대응하는 정도(正道)라는 생각이 듭니다. 다시 한 번 훌륭한 발제를 해주신 교수님께 감사드립니다. 〈끝〉

Session II

Technological Advancements, Crime, and Response of Criminal Justic

(첨단기술의 발전과 범죄, 그리고 형사사법의 대응)

Moderator: Hahn, Myung Kwan

Partner, Barun Law LLC/Advisor

Advisor, Fourth Industrial Revolution Convergence Law Association

사회: 한명관

법무법인 바른 변호사

제4차산업혁명융합법학회 고문

Session II

Technologies avancées et Justice pénale: les défis de la nouvelle criminalité et les promesses pour la justice pénale)

(첨단기술과 형사사법: 신종범죄의 도전과 형사사법의 전망)

Eric Mathais

State prosecutor

Bobigny Judicial Court

프랑스 보비니 검찰청 검사장

MINISTÈRE
DE LA JUSTICE
Liberté
Égalité
Fraternité

Cour d'appel de Paris
Tribunal judiciaire de Bobigny
Le procureur de la République

TRIBUNAL JUDICIAIRE
DE BOBIGNY

Eric MATHAIS
procureur de la République près le tribunal judiciaire de BOBIGNY

Technologies avancées et Justice pénale : les défis de la nouvelle criminalité et les promesses pour la justice pénale

1

1 ère partie : Technologies avancées et défis de la nouvelle criminalité

I. Un phénomène criminel en mutation

1. Emergence de la cybercriminalité
2. Ampleur du phénomène

II. Adaptation normative et institutionnelle

1. Évolution du cadre légal
2. Spécialisation des acteurs policiers et judiciaires

III. Les défis de la réponse judiciaire

1. Les difficultés d'enquête et de preuve
2. Vers une justice numériques et proactive

2

Seconde partie : Technologies avancées au service de la justice pénale

I. Considérations générales

1. Intelligence artificielle et traitement massif des données
2. Dématérialisation et accessibilité de la justice
3. Impact des nouvelles technologies sur notre façon de travailler

II. Panorama des nouvelles technologies utilisées dans la justice pénale française

1. Applicatif Cassiopée
2. Casier judiciaire européen numérisé
3. Outils informatiques modernes pour les procureurs et la Justice pénale
4. Procédure pénale numérique

III. Les risques et les défis éthiques

1. Protection des droits fondamentaux et de la vie privée
2. Souveraineté numérique
3. Transparence et supervision humaine

3

Conclusion

Vers une justice pénale augmentée mais humaine

- La technologie n'est pas une menace si elle reste au service de la justice
- Mais elle ne doit par remplacer la conscience
- Une justice éclairé par la donnée informatique, mais guidée par l'humanité

4

Merci de votre attention

Eric MATHAIS,

Procureur de la République près le Tribunal judiciaire de Bobigny

MINISTÈRE
DE LA JUSTICE
Liberté
Égalité
Fraternité

Cour d'appel de Paris
Tribunal judiciaire de Bobigny
Le procureur de la République

TRIBUNAL JUDICIAIRE
DE BOBIGNY

Eric MATHAIS
프랑스 보비니 검찰청 검사장

첨단기술과 형사사법:
신종범죄의 도전과 형사사법의 전망

제1부: 첨단기술과 신종범죄의 도전

I. 변화하는 범죄 현상
1. 사이버범죄의 등장
2. 현상의 규모

II. 규범적•제도적 적응
1. 법적 틀의 진전
2. 경찰 및 사법기관 담당자들의 전문화

III. 사법적 대응의 도전과제
1. 수사 및 증거 확보의 어려움
2. 선제적이며 디지털화된 사법을 향해

제2부: 형사사법에 활용되는 첨단기술

I. 일반적 고찰(고려사항)

1. 인공지능과 빅데이터 처리
2. 사법의 전자화(Dématérialisation)와 접근성
3. 새로운 기술이 우리의 업무방식에 미치는 영향

II. 현재 프랑스 형사사법에서 활용 중인 새로운 기술들의 개요

1. 카시오페(Cassiopée) 애플리케이션
2. 디지털화된 유럽 전과기록(범죄경력조회) 시스템
3. 검찰과 형사사법 제도를 위한 현대적 정보처리 수단
4. 디지털 형사절차(Procédure pénale numérique)

III. 위험과 윤리적 과제

1. 기본권과 사생활(프라이버시) 보호
2. 디지털 주권(Souveraineté numérique)
3. 투명성과 인간에 의한 감독

3

결론

인간적이면서도 증강된 형사사법제도를 향하여

- 사법을 위한 도구로 남아 있는 한, 기술은 위협이 되지 않습니다.
- 그러나 기술이 의식/양심(conscience)을 대체해서는 안 됩니다.
- 그리고 데이터에 의해 명약관화해지되, 인간성에 의해 인도되는 사법을 지향해야 합니다.

4

경청해 주셔서 감사합니다.

Eric MATHAIS,

프랑스 보비니 검찰청 검사장

MINISTÈRE
DE LA JUSTICE
Liberté
Égalité
Fraternité

Cour d'appel de Paris
Tribunal judiciaire de Bobigny
Le procureur de la République

TRIBUNAL JUDICIAIRE
DE BOBIGNY

SEOUL
International Seminar on Advanced Technologies and Criminal Justice

Technological Advancements, Crime, and Response of Criminal Justice
Presentation by Eric MATHAIS (State prosecutor, BOBIGNY Judicial Court)

***

Technologies avancées et Justice pénale : les défis de la nouvelle criminalité et les promesses pour la justice pénale

Introduction

Mesdames et Messieurs, Chers collègues,

C'est avec une profonde gratitude et un très grand intérêt que je prends la parole aujourd'hui à Séoul, lors de ce passionnant séminaire international consacré aux technologies avancées et à la justice pénale.

En tant que procureur de la République, je mesure chaque jour combien l'innovation technologique bouleverse notre rapport, au temps, à l'espace, à la preuve et, plus largement, à la vérité judiciaire.

Elle bouleverse même notre façon de travailler.

Nous assistons, depuis une ou deux décennies, à une mutation accélérée :

- l'intelligence artificielle,
- la science des données,
- la reconnaissance faciale,
- les outils de prédiction,
- la blockchain
- et les environnements numériques sécurisés (du moins nous l'espérons…),

redessinent le paysage de la justice.

Mais à cette promesse d'efficacité s'ajoute un vertige éthique.

Le risque est que la technologie, si elle n'est pas maîtrisée, se substitue peu à peu à l'humain, à son discernement, à sa capacité d'interprétation.

Il existe aussi un enjeu majeur de protection et de souveraineté nationale des données.

D'autre part, il existe une autre face du progrès technologique : le progrès technologique au service des malfaiteurs.

Les délinquants savent très bien utiliser les nouvelles technologies pour commettre des infractions et pour faire disparaître les preuves ou recycler l'argent sale de la criminalité.

Comment la justice pénale s'adapte-t-elle ?

Je commencerai mon propos par cette face-là du progrès technologique.

Et puis j'aborderai la question de la technologie avancée au service de la justice.

Première partie : Technologies avancées et défis de la nouvelle criminalité

I. Un phénomène criminel en mutation

1. L'émergence de la cybercriminalité

Les nouvelles technologies ont permis la naissance de formes inédites d'infractions.

Des atteintes aux systèmes de traitement automatisé de données :

- piratage,
- introduction frauduleuse dans un système informatique,
- altération de données, etc.

Mais aussi des infractions classiques commises par des moyens numériques :

- escroqueries en ligne,
- diffamation ou injure sur internet,
- harcèlement en ligne,
- usurpation d'identité numérique,
- atteintes à la vie privée,
- diffusion d'images non consenties, etc.

Enfin de nouvelles zones criminogènes ont vu le jour :

- dark-web,
- cryptoactifs,
- intelligence artificielle (deepfakes, fraudes automatisées)…

2. L'ampleur du phénomène

En 2024, on estime en France que près d'un tiers des faits de délinquance économique et financière signalés, avaient un volet numérique (source : ONDRP / ministère de l'Intérieur).

II. Une adaptation normative et institutionnelle

La justice et la Loi françaises se sont adaptées progressivement.

1. L'évolution du cadre légal

A la fin des années 80 sont apparues premières incriminations des atteintes aux services de traitement automatisé des données.

Loi du 21 juin 2004 pour la confiance dans l'économie numérique (LCEN) : possibilité de mettre en jeu la responsabilité pénale des hébergeurs et fournisseurs d'accès.

Loi du 13 novembre 2014 et loi du 24 août 2021 (loi sur le séparatisme, essentiellement religieux) : renforcement des pouvoirs d'enquête et de suppression de contenus terroristes ou haineux en ligne.

Règlement européen DSA (Digital Services Act) et DMA (Digital Markets Act) : encadrement des grandes plateformes (entrée en application 2024–2025).

2. La spécialisation des acteurs policiers et judiciaires

Une technicité particulière est désormais nécessaire pour traiter ce type de délinquance.

- ✓ 9 juridictions interrégionales spécialisées (JIRS) ont été mises en place en 2004. Elles sont compétentes pour **juger deux types d'infractions, quand elles sont d'une grande complexité** et **notamment en matière de cybercriminalité** :
  - **le crime organisé**
  - **la délinquance financière**

- ✓ Une juridiction nationale de lutte contre la criminalité organisée (JUNALCO) a été créée en 2019 au sein du tribunal judiciaire de Paris. Elle est **en charge de la lutte contre la criminalité de très grande complexité dans les mêmes domaines que les juridictions interrégionales spécialisées** mais à un niveau supérieur de complexité ou de gravité.

  Dans cette JUNALCO, des procureurs sont spécialisés en cybercriminalité, avec une compétence possible sur tout le territoire pour les affaires complexes. Des juges d'instructions du tribunal de Paris sont spécialisés de la même façon.

- ✓ Afin de **lutter contre les propos haineux ou discriminatoires et les appels à la violence dans l'espace public numérique**, le **pôle national de lutte contre la haine en ligne** a été créé en 2020 au sein du tribunal judiciaire de Paris.

- ✓ En 2026 l'organisation judiciaire sera encore modifiée avec la création récente dans la loi, du PNACO, **procureur national anti criminalité organisée**.

Les procureurs peuvent avoir l'appui technique de **l'Office anti-cybercriminalité (OFAC)** un service d'enquête national qui a été créé en 2023 et qui a remplacé divers autres services.

Les principaux domaines d'intervention de ce service sont :
- la lutte contre les atteintes aux systèmes de traitement automatisé de données (les attaques informatiques et rançongiciels notamment) ;
- la lutte contre les cyber services criminels ;
- l'accompagnement dans la mise en œuvre de cyber-investigations ;
- les contenus illicites sur l'Internet.

Par ailleurs, de plus en plus d'enquêteurs peuvent réaliser des enquêtes sous pseudonyme sur Internet, dites « cyber-patrouilles ».

En parallèle, le ministère de la justice a développé un **réseau de procureurs référents** « nouvelles technologies » dans chaque tribunal.

Et il est possible de faire appel en la matière à une coopération accrue avec Europol et Eurojust pour les dossiers transfrontaliers.

III. Les défis de la réponse judiciaire

1. Les difficultés d'enquête et de preuve

Les enquêtes et les dossiers pénaux doivent pour pouvoir être jugés relever de nombreux défis ou difficulté :

- Anonymat des auteurs, usage de virtuals pivates networks ou de réseaux chiffrés.
- Localisation des serveurs à l'étranger et multiplicité des législations.
- Preuve numérique fragile (chaîne de conservation, intégrité des données).
- Nécessité d'une expertise technique pointue pour exploiter les traces numériques.

2. Vers une justice numérique et proactive

Pour répondre à ces défis, il faut faire preuve d'imagination, d'innovation et d'agilité numérique

Les plates-formes de signalement en ligne de ces infractions se sont beaucoup développées ces dernières années.

Ainsi par exemple la plate-forme **PHAROS**, acronyme de « Plateforme d'harmonisation, d'analyse, de recoupement et d'orientation des signalements », est une plateforme gouvernementale française de signalement des contenus et comportements en ligne illicites.

La plateforme, créée en 2009, est mise en œuvre l'Office anti-cybercriminalité (OFAC).

Elle est constituée d'une d'environ 50 enquêteurs

Plusieurs millions de signalements ont été traités depuis sa création, avec plus de 4.400 signalements hebdomadaires, en moyenne, dont 57 % concernent des escroqueries ou infractions financières.

Les agents de la plateforme filtrent les messages les plus inquiétants pour les transmettre en urgence aux services de police, de renseignement ou antiterroristes (dont environ 330 par an en urgence absolue à des fins de protection des victimes).

Pharos, en 2024, a demandé aux hébergeurs le retrait de 67.300 contenus classés « atteintes sexuelles sur mineurs », 2.600 contenus liés au terrorisme et 1.200 contenus discriminatoires, haineux.

En 2022 a été créé **THESEE**, (acronyme de Traitement Harmonisé des Enquêtes et Signalements pour les E-Escroqueries), la première plate-forme de signalement et de plainte en ligne du ministère de l'Intérieur pour les escroqueries sur internet.

Les internautes peuvent déposer une plainte en ligne en remplissant un formulaire personnalisé, qui sera validé puis signé électroniquement par un policier.

À terme, l'intelligence artificielle sera utilisée pour la détection de fraudes et l'analyse de masse de données.

Conclusion

La justice française, longtemps en retard, s'est dotée d'un arsenal législatif et institutionnel solide pour faire face aux infractions liées aux nouvelles technologies.

Mais la course reste asymétrique : les innovations technologiques précèdent souvent la réponse judiciaire.

Le défi majeur réside désormais dans la formation, la coopération internationale, et la capacité d'adaptation continue des acteurs de la justice à un environnement numérique en mutation permanente.

Je voudrais maintenant souligner tout ce que les technologies avancées peuvent apporter à la justice pénale.

Seconde partie : les technologies avancées au service de la justice pénale

I- Considérations générales

Permettez-moi de vous livrer quelques considérations générales, avant de vous dresser un rapide panorama de l'état d'avancement de l'utilisation pratique des technologies avancées dans la Justice pénale en France.

1. L'intelligence artificielle et le traitement massif des données

L'intelligence artificielle a déjà pénétré l'univers judiciaire.

En France, le ministère de la Justice expérimente depuis plusieurs années des outils de traitement automatisé des dossiers, d'analyse de jurisprudence et de détection des fraudes.

Ces expérimentations soulignent aussi l'importance de la supervision humaine.

Un algorithme n'explique pas ses choix.

Il ne raisonne pas, il calcule.

La justice, elle, raisonne.

2. La dématérialisation et l'accessibilité de la justice

La crise sanitaire de 2020 a accéléré la dématérialisation.

En France, le projet « Portalis » vise à offrir au citoyen un guichet numérique unique.

A ma connaissance, la Corée du Sud a développé, un système d'e-justice intégral où chaque citoyen peut consulter ses procédures en ligne.

Ces évolutions favorisent la transparence et l'accessibilité, tout en permettant une gestion plus rationnelle des flux judiciaires.

La justice pénale devient ainsi plus proche, plus lisible, plus immédiate.

3. L'impact des nouvelles technologies sur notre façon de travailler

Il est important d'avoir conscience que les nouvelles technologies font profondément évoluer notre façon de travailler et même de structurer nos dossiers.

Le plus souvent ce sont des conséquences positives.

Et personne n'imaginerait vouloir revenir sur notre façon de travailler d'il y a 20 ou 30 ans, tant les avancées sont indéniables.

Mais, peut y avoir des conséquences moins positives et il faut en avoir conscience.

Un exemple très pratique, celui de la synthèse des dossiers.

Dans de nombreux cas, le procureur doit rédiger une synthèse écrite de dossiers.

Lorsque j'ai commencé mes premières fonctions dans la magistrature, en février 1990, les procureurs n'avaient pas d'ordinateurs.

Nous lisions les dossiers en version papier et nous rédigions au stylo ces synthèses.

Pour la rédiger il fallait lire d'abord l'intégralité des dossiers.

Depuis que nous avons tous des ordinateurs, nous prenons des notes informatiques au fil de notre lecture, pour rédiger notre synthèse.

Le résultat est tout à fait différent de ce que nous faisions sur papier.

Les synthèses sont beaucoup plus longues, car elles sont plus des résumés de dossiers que des synthèses véritables.

Cette évolution n'est pas positive, mais l'impact de l'outil informatique est tel qu'il a été au fil des années impossible de corriger cette évolution parfaitement identifiée !

Demain l'Intelligence artificielle va permettre de modifier encore nos pratiques.

II- Panorama sommaire des nouvelles technologies actuellement utilisées dans la Justice pénale en France

Je souhaiterais concrètement faire ici un rapide panorama des nouvelles technologies utilisées en France.

Je me concentrerai sur les outils plus spécifiquement judiciaires.

1- Évidemment, depuis plusieurs années nous disposons **d'un applicatif national « Cassiopée »,** recensant l'intégralité des procédures pénales enregistrées dans chaque tribunal, avec les décisions prises dans cette procédure.

Cela permet lorsqu'une personne est mise en cause à Bobigny de savoir en quelques secondes, si elle a déjà été mise en cause dans n'importe quel autre tribunal.

Cet outil recense toutes les procédures y compris les procédures n'ayant pas eu de suite.

Cela permet éventuellement de les ressortir au regard d'une nouvelle procédure.

À noter que désormais, les procédures des services d'enquête alimentent automatiquement l'applicatif informatique, ce qui fait gagner du temps d'enregistrement.

2- Nous disposons ensuite d'un **casier judiciaire national intégralement numérisé,** sur lequel figurent toutes les condamnations prononcées (beaucoup moins nombreuses donc que les procédures recensées par Cassiopée).

D'une part chaque nouvelle condamnation alimente de manière dématérialisée le casier judiciaire.

D'autre part depuis quelques années, ce casier judiciaire est interconnecté avec un nombre de plus en plus important de pays européens.

Il est possible en quelques secondes d'obtenir le casier judiciaire européen d'une personne.

3- Une des activités principales des procureurs français est la **permanence pénale**.

Il s'agit, 24 heures sur 24 possiblement, d'être en mesure d'échanger avec les enquêteurs pour :
- diriger les enquêtes pénales,
- ordonner des investigations à charge et à décharge,
- estimer à l'issue de l'enquête quelle orientation, quelle décision doit être prise
- et éventuellement saisir en urgence un juge d'instruction ou un tribunal pour faire juger un dossier à l'issue de la garde à vue du suspect.

A Bobigny, chaque 24 heures, les permanences traitent entre 500 et 600 appels téléphoniques, sans compter les mails et des centaines de dossiers.

Il est donc nécessaire d'avoir toute une batterie d'outils informatiques qui permettent la bonne gestion de ces permanences.

La quasi-totalité de ce travail se fait informatiquement à travers différents « **Outils informatiques modernes** ».

Chaque dossier fait l'objet de notes informatiques prises par le procureur qui gère le dossier. Je peux accéder éventuellement à chacun de ces dossiers pour savoir de quoi il s'agit et quelles sont les décisions qui ont été prises par les collègues.

4- Je souhaite enfin insister sur le **projet majeur du ministère de la justice qui va aboutir à terme à une dématérialisation complète des procédures pénales : le projet de procédure pénale numérique « PPN ».**

Le programme PPN constitue l'une des priorités du plan de transformation numérique de la Justice pénale afin de la rendre plus moderne, efficace et accessible au bénéfice des justiciables, des services enquêteurs et des juridictions.

L'objectif ultime est une transmission numérique des procédures pénales, une réception instantanée et une disparition totale des dossiers physiques.

- **4 Objectifs concrets de PPN :**
  1- Un travail collaboratif permettant à différents acteurs d'accéder simultanément au dossier pénal dématérialisé ;
  2- Une apposition de signatures sous format numérique
  3- De nouvelles fonctionnalités pour préparer les audiences : schématisation, recherche rapide et ajout de pièces multimédia au dossier pénal numérique.
  4- Une dématérialisation totale de la chaîne pénale, avec une disparition totale des dossiers en format papier.

III. Les risques et les défis éthiques

1. La protection des droits fondamentaux et de la vie privée

Toute innovation technologique en matière judiciaire doit se mesurer à l'aune de la protection des droits.

2. La souveraineté numérique

C'est un enjeu majeur.

Les logiciels d'IA, les plateformes d'hébergement, les infrastructures de cloud sont souvent d'origine extra-européenne.

Cela pose des questions de confidentialité, mais aussi d'indépendance de la puissance publique.

La France et l'Union européenne doivent renforcer leur capacité technologique autonome — c'est le sens des programmes GAIA-X et France Numérique 2030.

3. Transparence et supervision humaine

Toute décision judiciaire assistée par un algorithme doit être explicable.

Les magistrats doivent pouvoir comprendre et contrôler les critères retenus.

En 2018, une Charte éthique européenne sur l'utilisation de l'intelligence artificielle dans les systèmes judiciaires a été adoptée. Elle est fondée sur cinq principes :

- respect des droits fondamentaux,
- non-discrimination,
- qualité et sécurité,
- transparence
- et supervision humaine.

Conlusion : vers une justice pénale augmentée, mais humaine

La technologie n'est pas une menace si elle reste au service de la justice.

Elle peut rendre la procédure plus efficace, l'enquête plus rigoureuse, la preuve plus solide.

Mais elle ne doit jamais remplacer la conscience.

Comme le disait le juriste Français Jean Carbonnier, « le droit ne vaut que par les hommes qui l'appliquent ».

Nous devons concevoir une justice augmentée, non pas une justice automatisée.

Une justice éclairée par la donnée, mais guidée par l'humanité.

Il nous appartient de décider jusqu'où l'automatisation peut aller sans trahir l'esprit du droit.

En France, comme en Corée, comme partout où la démocratie repose sur la raison, nous devons affirmer que le procureur et le juge restent les garants ultimes de la Loi et du sens.

Les algorithmes calculent, les magistrats jugent.

Et c'est dans cette différence que réside la dignité de la justice humaine.

MINISTÈRE
DE LA JUSTICE
Liberté
Égalité
Fraternité

Cour d'appel de Paris
Tribunal judiciaire de Bobigny
Le procureur de la République

TRIBUNAL JUDICIAIRE
DE BOBIGNY

첨단기술과 형사법 국제세미나

첨단기술의 발전과 범죄, 그리고 형사사법의 대응
(Technological Advancements, Crime,
and Response of Criminal Justice)

Eric MATHAIS
(프랑스 보비니 검찰청 검사장)

첨단기술과 형사사법
(Technologies avancées et Justice pénale):
신종범죄의 도전과 형사사법의 전망
(les défis de la nouvelle criminalité
et les promesses pour la justice pénale)

서론

존경하는 동료 여러분, 오늘 서울에서 열리는 첨단기술과 형사사법을 주제로 한 이 흥미로운 국제세미나에서 여러분 앞에 서서 발표하게 된 데 대하여 깊은 감사와 큰 관심을 표하는 바입니다.

공화국 검사로서 저는 매일 같이 기술 혁신이 시간·공간·증거, 더 넓게는 사법적 진실에 대한 우리의 태도/방식을 얼마나 뒤흔들고 있는지 절감하고 있습니다.

기술 혁신은 심지어 우리의 업무 방식마저도 뒤흔들고 있습니다.

지난 10~20년 사이 우리는 다음과 같은 기술 혁신이 가져오는 급격한 변화를 목도하고 있

습니다.
- 인공지능(AI),
- 데이터 과학,
- 안면 인식,
- 예측 도구,
- 블록체인,
- 그리고 (적어도 우리는 그렇기를 기대하지만) 안전한 디지털 환경 등이 형사사법의 지형을 다시 그려 나가고 있습니다.

그러나 이러한 효율성에 대한 전망에는 윤리적 혼란 또한 수반됩니다.

기술이 통제되지 않을 경우, 기술은 인간의 판단력과 해석 능력을 점차 대체할 위험이 있습니다. 또한 데이터 보호와 데이터의 국가적 주권의 문제도 매우 중요합니다.

한편, 기술 발전에는 또 다른 측면이 존재하는데: 바로 범죄자들의 (범죄) 활동에 있어서도 기술 발전이 작용하고 있다는 점입니다.

범죄자들은 새로운 기술을 이용하여 능숙하게 범죄를 저지르고, 증거를 인멸하며, 범죄수익을 세탁하고 있습니다.

그렇다면 **형사사법은 이러한 변화에 어떻게 대응하고 있을까요?**

저는 우선 기술 발전의 이러한 측면에 대해 다루고, 그런 다음 **사법 업무에 활용되는 첨단 기술의 문제에 관해 논하고자 합니다.**

제1부 첨단기술과 신종범죄의 도전

Ⅰ. 변화하는 범죄 현상

1. 사이버범죄의 등장

새로운 기술은 과거에는 존재하지 않았던 새로운 유형의 범죄를 가져왔습니다.

○ 데이터 자동 처리 시스템에 대한 침해 범죄:
- 해킹
- 정보 시스템에 대한 무단 침입
- 데이터 변조 등

○ 디지털 수단을 통해 자행되는 전통적 범죄:
- 온라인 사기
- 인터넷상의 명예훼손 및 모욕
- 온라인 괴롭힘
- 디지털 신원 도용

- 사생활 침해
- 동의 없는 이미지 배포 등

○ 새로운 범죄 공간의 출현:
- 다크웹
- 암호화폐(cryptoactifs)
- 인공지능(딥페이크, 자동화된 사기 등)

2. 현상의 규모

2024년 프랑스에서 보고된 경제·금융범죄의 약 3분의 1에 디지털 요소가 포함된 것으로 추정됩니다(출처: ONDRP/내무부).

II. 규범적·제도적 적응

프랑스 사법과 법률은 점진적으로 변화에 대응해 왔습니다.

1. 법적 틀의 진전

○ 1980년대 후반: 데이터 자동 처리 시스템 침해행위를 처벌하는 규정이 최초로 등장

○ 2004년 6월 21일자 「디지털 경제 신뢰법(Loi pour la confiance dans l'économie numérique, LCEN)」: 웹호스팅 서비스 제공자 및 인터넷 서비스 제공업체에 대해 형사책임을 물을 수 있는 법적 근거 마련

○ 2014년 11월 13일자 및 2021년 8월 24일자 법률(분리주의에 관한 법률, 주로 종교적 분리주의에 대응): 온라인에서 테러·증오 콘텐츠를 조사하고 삭제할 수 있는 권한 강화

○ EU 디지털서비스법(DSA) 및 디지털시장법(DMA) 규정: 대형 플랫폼 규제(2024-2025년 발효)

2. 경찰 및 사법기관 담당자들의 전문화

이러한 유형의 범죄를 다루기 위해서는 이제 고도의 기술적 전문성을 필요로 합니다.

○ 2004년에 9개의 지역 간 전문법원(Juridictions interrégionales spécialisées, JIRS)이 설치되었습니다. 이 지역 간 전문법원은 **고도의 복잡성을 띠는 - 특히 사이버범죄와 관련된 -**

다음과 같은 2가지 유형의 범죄 사건들을 관할합니다:

- **조직범죄**
- **금융범죄**

○ 2019년, 파리 사법법원 내에 조직범죄 대응을 위한 국가법원(juridiction nationale de lutte contre la criminalité organisée, JUNALCO)이 신설되었습니다.

이 기관은 **지역 간 전문법원(JIRS)과 동일한 분야에서 매우 복잡한 범죄를 관할**하되, 더 높은 수준의 복잡성 또는 중대성을 지닌 사건을 담당합니다.

조직범죄 대응을 위한 국가법원(JUNALCO)에는 사이버범죄에 전문화된 검사가 배치되어 있으며, 복잡한 사건에 대해서는 전국적 관할권을 행사할 수 있습니다. 파리 사법법원 소속의 수사판사(juges d'instruction) 역시 마찬가지로 사이버범죄에 특화되어 있습니다.

○ 2020년, **디지털 공공 영역에서의 증오·차별 발언 및 폭력 선동에 대응하기 위해 "온라인 혐오 대응 국가전담부서(pôle national de lutte contre la haine en ligne)"**가 파리 사법법원 내에 설치되었습니다.

○ 2026년, **"조직범죄 대응 국가 검사(procureur national anti criminalité organisée, PNACO)"**가 최근 법률로 창설됨에 따라 사법 조직이 다시 한 번 개편될 예정입니다.

검사들은 2023년에 설립되어 여러 다른 부서들을 대체한 국가 수사기관인 **"사이버범죄국(Office anti-cybercriminalité, OFAC)"**의 기술적 지원을 받을 수 있습니다.

사이버범죄국의 주요 개입 분야는 다음과 같습니다.

- 데이터 자동 처리 시스템에 대한 침해 대응(특히 사이버공격 및 랜섬웨어)
- 범죄성 사이버 서비스(cyber services criminels) 대응
- 사이버 수사 수행을 위한 지원
- 인터넷 상의 불법 콘텐츠 대응

또한, 점점 더 많은 수사관들이 인터넷상에서 가명으로 수사를 수행할 수 있게 되었는데, 이를 '사이버 순찰(cyber-patrouilles)'이라 합니다.

이와 더불어, 법무부는 각 법원별로 '신기술(nouvelles technologies)' **담당 검사 네트워크(réseau de procureurs référents)**를 구축했습니다.

그리고 국경을 넘는 사건의 경우, 유로폴(Europol) 및 유로저스트(Eurojust)와의 협력 강화를 촉구하는 것도 가능해졌습니다.

III. 사법적 대응의 도전과제

1. 수사 및 증거 확보의 어려움

수사와 형사 사건이 재판에 회부되기 위해서는 다음과 같은 다양한 도전과 어려움을 극복해야 합니다.

- 범죄자의 익명성, VPN 및 암호화된 네트워크 사용
- 해외 서버 위치 및 국가별 다양한 법제도
- 디지털 증거의 취약성(데이터 보존 체인, 데이터 무결성 문제)
- 디지털 흔적을 분석하기 위한 고도의 기술적 전문성 필요

2. 선제적이며 디지털화된 사법을 향해

이러한 도전에 대응하기 위해서는 창의성과 혁신, 그리고 디지털 민첩성을 발휘해야 합니다.

최근 몇 년 동안 온라인 범죄 신고 플랫폼이 크게 발전했습니다.

그 예로 **PHAROS** 플랫폼을 들 수 있는데, "신고에 대한 조정, 분석, 교차검증 및 방침 안내를 위한 플랫폼(Plateforme d'harmonisation, d'analyse, de recoupement et d'orientation des signalements)"의 약자인 PHAROS 플랫폼은 불법 온라인 콘텐츠 및 행위를 신고하기 위한 프랑스 정부 플랫폼입니다.

2009년에 설립된 PHAROS 플랫폼은 사이버범죄국(Office anti-cybercriminalité, OFAC)에서 운영하고 있습니다.

PHAROS 플랫폼은 약 50명의 수사관으로 구성되어 있습니다.

PHAROS 플랫폼 창설 이래, 수백만 건의 신고가 처리되었는데, 매주 평균 4,400건의 신고가 접수되고 있으며, 이 중 57%가 사기 및 금융범죄와 관련되어 있습니다.

PHAROS 플랫폼 직원들은 가장 우려되는 메시지를 선별하여 경찰, 정보기관, 테러대응 부서 등에 긴급 전달합니다(피해자 보호 목적으로 연간 약 330건의 절대적 긴급(urgence absolue) 조치가 취해짐).

2024년, PHAROS 플랫폼은 "미성년에 대한 성적 착취(atteintes sexuelles sur mineurs)"로 분류된 콘텐츠 67,300건, 테러리즘 관련 콘텐츠 2,600건 그리고 차별/혐오 콘텐츠 1,200건에 대해 삭제를 요청한 바 있습니다.

2022년에는 "전자사기 사건의 수사 및 신고에 관한 조정 처리(Traitement Harmonisé des Enquêtes et Signalements pour les E-Escroqueries)"의 약자인 **THESEE**가 신설되었는데, THESEE는 인터넷 상의 사기에 대응하기 위해 내무부가 최초로 운영하는 첫 번째 온라인 신고 및 고소 플랫폼입니다.

인터넷 이용자는 맞춤형 양식을 작성해서 온라인으로 고소를 제기할 수 있는데, 이는 경찰관의 확인을 거쳐 전자 서명을 받게 됩니다.

향후에는 인공지능(AI)이 온라인 사기 탐지 및 빅데이터 분석에 활용될 것입니다.

소결

오랫동안 뒤처져 있던 프랑스 사법제도는 이제 새로운 기술과 관련된 범죄에 대응하기 위해 탄탄한 법적·제도적 장치를 갖추게 되었습니다.
그러나 이 흐름은 여전히 비대칭적이라 하겠는데, 많은 경우에 기술 혁신이 사법적 대응보다 앞서기 때문입니다.
앞으로의 핵심 과제는 교육/연수, 국제협력, 끊임없이 변화하는 디지털 환경에 대한 사법 행위자들의 지속적인 적응 능력에 있다 할 것입니다.

이제 저는 첨단기술이 형사사법에 어떤 기여를 할 수 있는지에 대해 말씀드리고자 합니다.

제2부 형사사법에 활용되는 첨단기술

Ⅰ. 일반적 고찰(고려사항)

프랑스 형사사법에서 첨단기술이 실제로 어떻게 활용되고 있는지 간략하게 현황을 말씀드리기에 앞서, 몇 가지 일반적 고찰을 공유하고자 합니다.

1. 인공지능과 빅데이터 처리

인공지능은 이미 사법 영역에 들어와 있습니다.
프랑스 법무부는 여러 해 전부터 사건 기록의 자동 처리, 판례 분석, 사기 탐지 도구를 실험해 왔습니다.
이러한 실험들은 인간에 의한 감독/통제의 중요성 또한 강조합니다.
알고리즘은 자신의 선택에 대해 설명하지 않습니다.
알고리즘은 사고(思考)/추론하지 않고, 계산합니다.
그러나 사법은 사고(思考)/추론합니다.

2. 사법의 전자화(dématérialisation)와 접근성

2020년 보건 위기는 전자화를 가속했습니다.
프랑스에서는 '포르탈리스(Portalis)' 프로젝트를 통해 국민에게 단일 디지털 창구를 제공하는 것을 목표로 하고 있습니다.

제가 알기로 대한민국은 모든 국민이 자신의 사건 절차를 온라인에서 조회할 수 있는 완전한 전자사법(e-Justice) 시스템을 구축하고 있습니다.
이러한 발전은 투명성과 접근성을 증진하는 동시에, 사법 흐름의 보다 합리적인 관리를 가능하게 합니다.
그 결과 형사사법은 더 가까워지고, 더 이해하기 쉬우며, 더 즉각적이 되어 가고 있습니다.

3. 새로운 기술이 우리의 업무방식에 미치는 영향

새로운 기술은 우리의 업무방식, 나아가 사건 기록을 구성하는 방식까지 깊이 변화시키고 있습니다.
대부분의 경우 그 결과가 긍정적이었던바, 20~30년 전의 업무방식으로 돌아가야 한다고 생각하는 사람은 아무도 없으리만큼, (기술적) 발전은 부인할 수 없습니다.
그러나 덜 긍정적인 결과도 있을 수 있으므로 이에 대한 인식이 필요합니다.
대표적인 실무 사례로 사건 기록의 요약(synthèse) 문제가 있습니다.
많은 경우에, 검사는 사건 기록에 대한 서면 요약을 작성해야 합니다.
제가 1990년 2월 법조계에서 처음 업무를 시작했을 당시, 검사들은 컴퓨터가 없었습니다.
우리는 사건 기록을 종이 문서로 읽고, 펜으로 요약을 직접 작성했습니다.
요약을 작성하기 위해서는 먼저 사건 기록 전체를 읽어야 했습니다.
하지만 오늘날 우리 검사 모두가 컴퓨터를 갖게 되자, 요약을 작성하기 위해 사건 기록을 읽어가며 그때그때 전자적 메모를 남기게 되었습니다.
그 결과는 종이로 작업하던 때와는 사뭇 다릅니다.
요약은 훨씬 더 길어졌는데, 진정한 '요약'이라기보다 사건 기록의 축약본(개요)에 가까워졌기 때문입니다.
이러한 변화를 긍정적이라고 보기 어렵지만, 정보처리기술(컴퓨터)의 영향력이 너무 큰 나머지, 수년에 걸쳐, 명확하게 확인된 이 발전을 되돌리는 것은 사실상 불가능했습니다.
앞으로 인공지능은 우리의 업무 관행에 다시 한 번 변화를 가져올 것입니다.

II. 현재 프랑스 형사사법에서 활용 중인 새로운 기술들의 개요

이제 프랑스에서 실제로 활용되고 있는 새로운 기술들에 대해 간단하게 소개하고자 합니다.
저는 특히 사법 영역에서 특별하게 사용되는 도구들에 중점을 두겠습니다.

1. 프랑스는 수년 전부터 **"카시오페(Cassiopée)"라는 국가 애플리케이션**을 사용하고 있는데, 카시오페(Cassiopée)는 각 법원에 등록된 형사절차와 그에 따른 결정들 전부를 목록화하고 있습니다.
카시오페(Cassiopée)는 보비니(Bobigny)에서 어떤 사건에 연루된 사람에 대해서 그가 다른 법원에서도 범죄에 연루된 적이 있는지 단 몇 초 만에 확인할 수 있도록 합니다.

카시오페(Cassiopée)는 후속 조치 없이 종결된 절차까지 포함하여 모든 절차를 기록합니다. 이를 통해 새로운 사건이 발생했을 때 과거 절차를 검색해서 참조할 수 있도록 합니다.
현재 수사기관의 절차가 자동으로 정보처리 애플리케이션에 연동되어 사건 등록 시간이 크게 단축되었다는 점에도 주목해야 합니다.

2. 다음으로, 프랑스는 모든 유죄판결이 기록된(카시오페(Cassiopée)에 기록된 절차보다 수적으로 훨씬 적음), **완전히 디지털화된 국가 전과기록(범죄경력조회) 시스템**을 갖추고 있습니다.
새로운 유죄 판결은 자동으로 전자 방식으로 전과기록 시스템에 등록됩니다.
최근 몇 년 전부터 이 전과기록 시스템은 점점 더 많은 유럽 국가와 상호 연결되고 있습니다.
이제는 한 사람의 유럽 전과기록을 불과 몇 초만에 조회하는 일이 가능하게 되었습니다.

3. 프랑스 검사의 주요 임무 중 하나는 **형사 당직(permanence pénale)** 업무입니다.
여기에는 필요할 경우 24시간 내내 수사관과 교신하며 다음과 같은 업무를 수행하는 것을 포함합니다.
- 형사 수사 지휘
- 피의자에게 유리하거나 불리한 수사 지시
- 수사 종료 시, 어떤 지침이나 결정을 내려야 할 것인지에 대한 판단
- 피의자 구금에 따라 사건을 재판에 부치기 위해, 필요한 경우 긴급하게 수사판사나 법원에 사건을 회부하는 것

보비니 검찰청에서는 당직 검사들이 하루 24시간 동안 500~600건의 전화 통화를 처리하며, 그밖에 이메일과 수백 건의 사건 기록을 다룹니다.
따라서 이러한 당직 업무가 원활히 운영되려면 다양한 정보처리 수단이 필수적입니다.
대부분의 업무는 여러 다양한 **"현대적 정보처리 수단(Outils informatiques modernes)"**을 통해 전자적 방식으로 처리됩니다.
각 사건마다 담당 검사가 전자적으로 메모를 작성합니다. 저 또한 필요한 경우 언제든지 사건 내용을 확인하고 동료들이 어떤 결정을 내렸는지 파악하기 위해 이러한 사건 기록들에 접근할 수 있습니다.

4. **프랑스 법무부에서 추진하고 있는, 궁극적으로 형사절차의 완전한 전자화로 귀결될 "디지털 형사절차(procédure pénale numérique, PPN)"** 프로젝트를 마지막으로 강조하고자 합니다.
PPN 프로젝트는 형사사법의 디지털 전환 계획의 우선 과제 중 하나로, (재판관할권 하에 있는) 소송당사자, 수사기관과 사법기관에게 이익이 되도록 형사사법을 보다 현대적이고, 효율적이며, 접근이 용이하게 만들고자 합니다.
(PPN 프로젝트의) 최종 목표는 형사절차의 완전한 디지털 전송, 즉각적 접수 및 물리적(종이) 사건 기록의 완전한 소멸입니다.

○ PPN 프로젝트의 4가지 구체적 목표는 다음과 같습니다.

1. 여러 사법 주체가 전자화된 형사기록에 동시에 접근할 수 있도록 하는 협업적 업무 환경 구축
2. 디지털 형식의 서명 첨부
3. 심문/공판 준비를 위한 새로운 기능: 도식화, 신속한 검색, 디지털화된 형사 기록에 멀티미디어 서류 첨부 등
4. 종이 형태의 사건 기록의 전면적 폐지와 더불어, 형사사법 전(全) 과정의 완전한 전자화

III. 위험과 윤리적 과제

1. 기본권과 사생활(프라이버시) 보호

사법 영역에서의 모든 기술 혁신은 기본권 보호의 기준에 따라 평가되어야 합니다.

2. 디지털 주권

디지털 주권(souveraineté numérique)은 매우 중요한 쟁점입니다.
많은 경우에, AI 소프트웨어, 호스팅 플랫폼, 클라우드 인프라 등은 유럽 외부에서 창안되었습니다.
이로 인해 기밀성 문제뿐만 아니라 공권력의 독립성 문제가 제기됩니다.
프랑스와 EU는 독자적인 기술 역량을 강화해야 하는바, 이는 가이아-X(GAIA-X)와 디지털 프랑스 2030(France Numérique 2030)과 같은 프로그램이 의도하는 방향이기도 하다.

3. 투명성과 인간에 의한 감독/통제

알고리즘의 도움을 받은 모든 사법적 결정은 설명 가능해야 합니다.
법관은 사용된 기준에 대해 이해하고 확인/통제할 수 있어야 합니다.
2018년에 사법 체계에서의 인공지능(AI) 사용에 관한 유럽 윤리헌장이 채택되었는바, 동 윤리헌장은 다음과 같은 다섯 가지 원칙에 기반합니다.

- 기본권 존중
- 차별 금지
- 품질과 보안
- 투명성
- 인간에 의한 감독/통제

결론: 인간적이면서도 증강된 형사사법을 향하여

사법을 위한 도구로 남아 있는 한, 기술은 위협이 아닙니다.

기술은 절차를 더 효율적으로 만들고, 수사를 더 엄밀하게 만들며, 증거를 더 확고하게 만들 수 있습니다.

그러나 기술이 의식/양심을 대체해서는 안 됩니다.

프랑스 법학자 장 카르보니에(Jean Carbonnier)가 말한 바와 같이, “법은 그것을 적용하는 사람들에 의해서만 가치가 있습니다(le droit ne vaut que par les hommes qui l’appliquent).”

우리가 지향해야 할 것은 자동화된 사법(justice automatisée)이 아니라 증강된 사법(justice augmentée)입니다.

데이터에 의해 명약관화해지되, 인간성에 의해 인도되는 사법.

법의 정신을 훼손하지 않는 한도에서, 우리는 자동화의 범위가 어디까지일지 결정해야 합니다.

프랑스에서도, 한국에서도, 그리고 이성에 기반한 민주주의가 존재하는 곳이라면 어디에서나 검사와 판사가 법과 그 의미를 지키는 최후의 보루로 남아야 함을 분명히 해야 합니다.

알고리즘은 계산하고, 법관은 판단합니다.

그리고 바로 이 차이 속에 인간적 사법의 존엄성이 존재합니다.

Session II

첨단기술을 이용한 범죄와 대응

Yun, Jee-Young

Senior Research Fellow,
Korean Institute of Criminology and Justice

윤 지 영

한국형사·법무정책연구원 선임연구위원

첨단기술을 이용한 범죄와 대응

윤 지 영
한국형사·법무정책연구원

1

Wassily Kandinsky, *The Blue Mountain*. 1908/09. / *Cemetery and Vicarage in Kochel*. 1909. / *Arabs I (Cemetry)*. 1909.

Old Town II. 1902. / *Beach Baskets in Holland*. 1904.

I. 들어가며

2

2025년 10대 미래 기술(MIT 테크놀로지 리뷰)

1. 소형언어모델
2. 베라 루빈 천문대
3. 장기지속형 HIV 예방제
4. 생성형 AI 검색
5. 소 트림 감소제
6. 청정 제트연료
7. 고속학습 로봇
8. 효과적인 줄기세포 치료
9. 로보택시
10. 녹색철강

3

Improvisation 19. 1911. / *Impression III (Concert).* 1911.

II. 첨단기술을 이용한 범죄

4

AI 2027 보고서

"'인공지능이 인류를 말살한다?' 큰 파장을 낳은 AI2027 보고서", BBC News 코리아, 2025. 8. 8, https://www.youtube.com/watch?v=sK9XnzP-YSQ

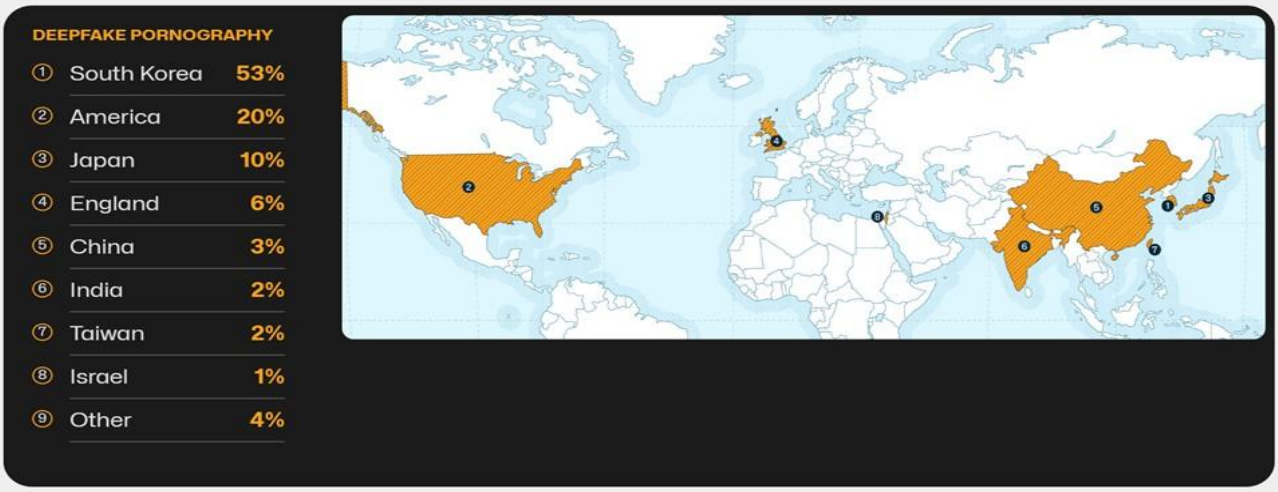
인공지능 이용 범죄

❖ 범죄의 지능화·고도화

- 딥페이크 이용 허위사실 유포, 피싱, 음란물 제작
- 금융이나 주식 시장 패턴 분석을 통한 시장 교란
- 소셜 미디어 정보 분석 등을 통한 맞춤형 피싱
- 취약점 분석 등을 통한 사이버 공격 성공 가능성 제고
- 생성형 인공지능 등장으로 인한 접근 용이성 제고

6

딥페이크 성착위물 피해자 국적



Security Hero 웹사이트, https://www.securityhero.io/state-of-deepfakes/

"유재석입니다"…유명인 사칭 '피싱' 판치는데 속수무책

피해자 대부분 6070·피해액 1조…해외 플랫폼에 집단소송 준비
AI 발달로 피해 확산 불 보듯…국제 공조·빅테크 핀셋 규제 필요

김은성, ""유재석입니다"…유명인 사칭 '피싱' 판치는데 속수무책", 주간경향, 2024. 4. 8,
https://weekly.khan.co.kr/khnm.html?mode=view&code=115&artid=202404010600041

경향신문

이러라고 만든 AI가 아닌데…챗GPT, 범죄조직 '사기 필수템' 오명

입력 2025.09.16 13:57
수정 2025.09.16 14:58

윤기은 기자 energyeun@kyunghyang.com

> 범죄 조직, 인신매매 피해자들 모아 강제 노역
> 사투리 흉내 내는 등 '스캠' 효율적 도구로 전락
> 오픈AI "오용 차단 노력…조사관들이 감시 중"

"이러라고 만든 AI가 아닌데…챗GPT, 범죄조직 '사기 필수템' 오명", 경향신문, 2025. 9. 16, https://www.khan.co.kr/article/202509161357001

"촬영물이야? AI 창작물이야?...오픈AI '소라' 공개", YTN, 2024. 2. 17, https://www.youtube.com/watch?v=AjOVV3sC0fc&t=72s

"실제 같은 AI 영상 확산…자칫하면 '독'", /SBS 8뉴스, 2025. 7. 27, https://www.youtube.com/watch?v=Q7KRc2TIrI8

"공장서 출고된 차가 스스로 차주 집으로…테슬라, '무인 배송' 첫선", MBN 뉴스, 2025. 6. 30, https://www.youtube.com/watch?v=IE41D0CcjWU

자율주행차 이용 범죄

❖ 테러 수단으로 활용

- **2016년 7월 14일** 프랑스 니스에서 대형트럭을 이용한 테러 발생
  - 이슬람 극단주의자로 추정되는 **30**대 남성이 대형 트럭을 몰고 군중에게 돌진하여 **84**명 사망, **100**여 명 부상 / 범인은 경찰과의 총격전 끝에 현장에서 사살
- 무인자동차가 테러에 이용될 경우 범죄자는 사상을 입을 염려가 없고, 발각될 가능성도 낮아짐

❖ 자율주행차 해킹을 통한 살인, 납치, 감금

❖ 자율주행차가 성매매 등 범죄 공간으로 활용

13

자율주행자동차에 대한 압수·수색

- 무인택시에 승객이 탑승한 경우, 그 차량에 대한 영장 제시 및 영장집행 사실 통보는 어떻게 진행?
- 수사기관은 관리자에게 영장집행 사실을 통보하고 참여하게 하는 절차 진행
- 승객이 탑승하고 있는 승객 칸이 수색의 대상이 된 때에는 탑승객에게 수색영장 제시
- 미국 판례에 의할 때, 자동차에 대한 합법적인 영장집행이 이루어지면 트렁크를 포함한 자동차의 모든 부분이 수색의 대상이 되고, 승객 칸은 물론이고 차 안에 있는 승객의 소지품도 수색의 대상

14

- 승객이 미처 알지 못하는 사이에 범인이 승객의 소지품에 금제품을 은닉할 수 있다는 가능성이 고려된 것이나, 차량에 대한 수색 영장만으로 승객에 대한 수색까지 이루어지는 것에 대한 비판 가능
- 주행 중인 무인 자율주행차의 경우 피처분자의 부재로 인하여 영장제시 없이 압수·수색 가능
- 영장제시의 실질적 의미는 영장의 내용과 범위를 피처분자가 인지하도록 하는 것이라는 점에 주목하며 효율적 방식 고려 필요
- **2021년 10월 19일** 도입된 전자영장이 이용될 경우 피처분자가 현장에 부재중인 상황에서도 원격지에서 영장을 제시할 수 있을 것인바, 압수·수색의 적법요건을 충족하면서 수사상의 목적이 달성될 수 있을 것으로 기대

15

자율주행자동차 운행 정보 취득

- 차량의 소유자뿐만 아니라 전기통신사업자, 위치정보사업자, 기반시설 운영 주체 및 차량 제조사가 차량의 운행 정보 보유 가능
- 수사기관이 「전기통신사업법」에 의해 요청할 수 있는 통신사실 확인자료에는 발신기지국의 위치추적 자료 포함
- **Carpenter v. United States(2018)** 판결 취지가 휴대전화가 아닌 자율주행차 위치정보에도 적용?

- 일반적으로 자동차는 기동성, 공공 도로상 규제 가능성, 프라이버시에 대한 낮은 기대 등을 이유로 영장 없는 수색이 광범위하게 인정
- 운행 중인 자율주행자동차에 대해서는 영장 없는 수색 가능성이 넓게 열릴 것

16

AI 드론

"'드론이 사람을 노리는 시대", 스브스뉴스, 2018. 8. 7, https://www.youtube.com/watch?v=JboHVB5ijQ0

뇌-컴퓨터 인터페이스(Brain-Computer Interface, BCI)

"뇌와 컴퓨터를 연결, 'BCI' 기술이란?", YTN, 2024. 1. 31, https://www.youtube.com/watch?v=Kop2cMgM3xw

"머스크 기술에 도전장 던졌다...'인간 뇌 칩' 불붙은 경쟁", YTN, 2024. 12. 18, https://www.youtube.com/watch?v=JsjYCv1sr-o

"'유전자 가위 크리스퍼 치료제' 첫 출시 임박!...새 시대 눈앞", YTN, 2023. 12. 7, https://www.youtube.com/watch?v=MjmJRtyzQs4

유전자 가위 혁명…결함 DNA 콕 집어 희소질환 아기 생명 구했다

송고 2025-05-16 08:39

생후 시술…노벨상 '크리스퍼' 기술 토대로 성공한 첫 사례
"수십년 약속 결실…세상이 의학 접근하는 방식 바꿔놓을 것"

유전자 치료를 받은 KJ [AP 연합뉴스, 필라델피아 아동병원 제공. 재판매 및 DB 금지]

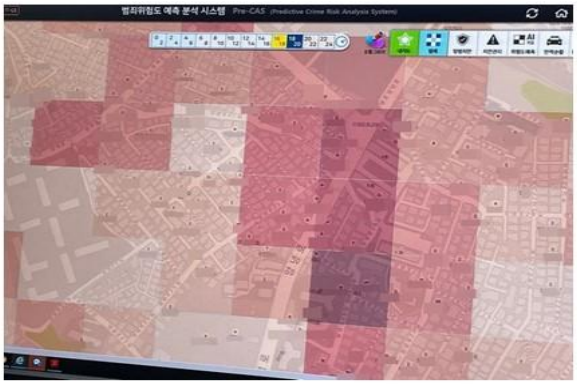
"유전자 가위 혁명…결함 DNA 콕 집어 희소질환 아기 생명 구했다", 연합뉴스, 2025. 5. 16, https://www.yna.co.kr/view/AKR20250516024500009?input=1195m

Old Town II. 1902. / *Beach Baskets in Holland.* 1904.

III. 첨단기술을 이용한 범죄 대응

22

인공지능 프리카스(pre-Cas)



"[경찰, 과학과 만나다] ① "여기서 우회전하세요"… 우범지역 찾아주는 AI, 순찰 풍경 바꿨다", 조선비즈, 2021. 10. 21, https://biz.chosun.com/topics/topics_social/2021/10/21/62YGGIBWLNFGBHZNK4R6X75FXA/

""실시간 번역과 사진을 동영상으로…" 구글, AI 신기술 '눈길' ",JTBC News, 2025. 8. 21, https://www.youtube.com/watch?v=VjpiK4ePE7c

순찰 로봇

(좌):"24시간 골목을 누비며 감시하는 '순찰 로봇' 등장", 매일경제 2014. 9. 1, https://www.mk.co.kr/news/photo/view/2014/09/1156219/
(우): 발표자 촬영 사진

25

로봇

- 미국의 나이트스코프(Knightscope)사가 개발한 'K5'
  - 사전에 입력된 경로를 따라 자율주행, 탑재된 360도 회전 카메라를 이용해 자동차 번호판 식별, 적외선 열화상 센서를 이용해 야간에도 사람이나 사물 인식, 인식 정보나 메시지 송·수신, 긴급 상황 발생 시 경찰 호출 가능
  - 2016년 7월 8일 미국 텍사스 주 댈러스에서는 경찰관을 저격한 범인을 사살하는 과정에 '폭탄 로봇(bomb robot)' 투입
  - 2016년 12월 8일 러시아는 폭발물을 탑재한 장갑차 모양의 킬러로봇을 이용하여 출입물을 폭파하고 내부로 진입하여 IS 조직원 사살

26

"논란의 '로봇 개' 결국 투입…"안전 위한 결정"", MBC NEWS, 2023. 4. 17, https://www.youtube.com/watch?v=gJFy_HepNm0&t=66s

Tesla Optimus Robot Explained, DPCcars, 2024. 10. 13, ttps://www.youtube.com/watch?v=1t1KMXSBrBM

"대화로 움직이는 로봇?…오픈 AI, 또 일냈다", JTBC News, 2024. 3. 14, https://www.youtube.com/watch?v=bAr75K-5Uqw

로봇 교도관

30

스마트 안경

(좌): "中国警察姐姐佩戴智能眼镜执法 帅炸了！", 每日头条, 2018.2.10., https://kknews.cc/zh-sg/news/x3zgeoo.html
(우): "阿根廷又被中国科技震惊了！警察配上智能眼镜两秒识别身份证", 搜狐, 2018. 2. 15., https://www.sohu.com/a/222859597_155500

31

- 스마트안경
- 중국이나 아랍에미리트에서는 증강현실(AR) 기술이 적용된 스마트안경을 이용하여 용의자 수색 및 수배 차량 식별
- **2020년 8월,** 우리 정부는 용의자 및 수배차량 조회에 이용될 수 있는 스마트안경 도입 방안을 모색하겠다고 밝히기도 함
- 수사기관이 스마트안경을 도입할 경우 어떤 기능을 탑재하는지에 따라서 관련 법규의 정비 범위 상이
- 수배차량의 조회 기능은 큰 논란 없이 활용될 수 있으나, 얼굴인식을 통한 용의자 신원 조회는 국민적 반발에 부딪혀 시행될 수 없을 것으로 전망

32

스마트 안경

"'실시간 번역부터 AI 대화까지' 구글 스마트 안경 공개", JTBC News, 2025. 5. 21., https://www.youtube.com/watch?v=rwvm-x_aZIE

Improvisation 19. 1911. / *Impression III (Concert).* 1911.

V. 나가며

첨단기술과 법의 관계

35

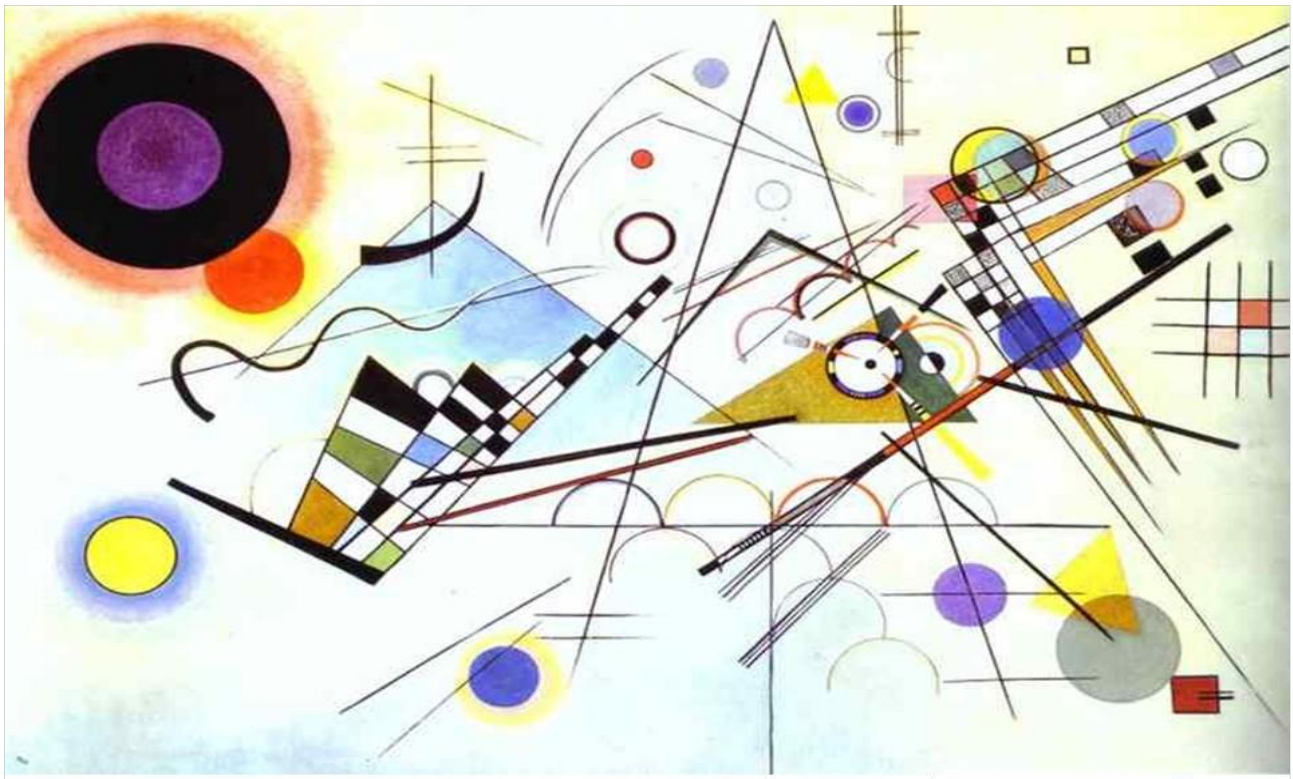
Easter morning 1900: 5th Ave, New York City. Spot the automobile.

Source: US National Archives.

Easter morning 1913: 5th Ave, New York City. Spot the horse.

Source: George Grantham Bain Collection.

36



Composition VIII. 1923.

THANK YOU

37

Discussion Paper on Session 2

Technological Advancements, Crime, and Response of Criminal Justice

Kim, Dae Won

Visiting Professor, Inha University Law School

Lee, Won-Sang

Professor, Department of law, Chosun University

Ryu, Bu-Gon

Professor, Department of law, Korean National Police University

토론문

[제2주제] 첨단기술의 발전과 범죄, 그리고 형사사법의 대응에 관한 토론문

김대원
(인하대학교 법학전문대학원 초빙교수)

먼저 첨단 기술이 프랑스 형사사법체계에 어떻게 사용되는지 자세한 설명을 해주신 Eric MATHAIS 검사장님께 감사의 말씀을 드립니다.

감사장님의 발표는 인공지능(AI)과 빅데이터의 활용, 사법 절차의 전자화 프로젝트(PPN) 등 형사사법에 도입되고 있는 첨단 기술들을 개괄하고 있으며, 이 과정에서 발생하는 윤리적 도전과 기본권 보호의 중요성을 강조합니다. 결론적으로, 기술은 사법 절차를 효율화하지만 인간적인 판단과 감독이 최종적으로 법의 존엄성을 지켜야 한다고 역설하시고 있는데 전적으로 동의하는 바입니다.

따라서 저는 검사장님의 발표에 토론이라기 보다는 프랑스제도에 대한 몇 가지 궁금한 점을 질문드리고자 합니다.

먼저, 발표문에 따르면 “Ⅱ. 규범적·제도적 적응”과 관련하여 2026년 “조직범죄 대응 국가 검사(procureur national anti criminalitéorganisée, PNACO)”가 최근 법률로 창설됨에 따라 사법 조직이 다시 한 번 개편될 것이라고 하셨습니다. 새롭게 도입되는 검사는 기존의 검사에 비하여 어떠한 권한이 더 부여되는지, 아니면 수사범위가 조정되는지 등 기존의 검사와 차이점이 무엇인지 설명을 부탁드립니다.

두 번째는, “점점 더 많은 수사관들이 인터넷상에서의 가명으로 수사를 수행할 수 있게 되었는데, 이를 ‘사이버 순찰(cyber-patrouilles)’이라 합니다.”라고 하셨습니다. 현재 한국에서는 이를 위장수사라고 부르고 있으며, 위장수사를 할 수 있는 범죄는 아동청소년

을 대상으로 한 사이버 성범죄에 국한되고 있습니다. 프랑스의 경우 '사이버 순찰(cyber-patrouilles)'의 범위와 성과(효과)에 대하여 추가 설명을 부탁드립니다.

세 번째는, 선제적이며 디지털화된 사법을 향해서 범죄신고 플랫폼인 'PHAROS 플랫폼'과 'THESEE 플랫폼'을 운영하고 있다고 하셨는데, 양자의 차이점이 무엇인지, 왜 별도로 운영하는지에 대한 추가 설명을 부탁드립니다.

마지막으로 프랑스 법무부가 추진하고 있는 "디지털 형사절차(procédure pénale numérique, PPN)" 프로젝트에 대한 것입니다. 프랑스 법무부는 이 프로젝트의 최종목표로 형사절차의 완전한 디지털 전송, 즉각적 접수 및 물리적(종이) 사건 기록의 완전한 소멸을 들고 있습니다.

현재 한국의 형사사법 전자화는 주로 다음 두 가지 법률을 근거로 추진되고 있습니다. 「형사사법절차 전자화 촉진법」은 형사사법절차의 전자화를 촉진하고, 기관 간 정보의 공동 활용 및 대국민 서비스 개선을 목적으로 합니다. 또한 「형사사법절차에서의 전자문서 이용 등에 관한 법률」은 형사사법절차에서 전자문서의 사용을 원칙으로 하고, 종이문서와 동일한 효력을 부여하는 법적 근거를 마련했습니다. 이 법률은 2024년부터 전면 시행되어 형사사법 절차의 "전자화(형사전자소송)"를 본격화하는 핵심 기반이 됩니다. 한국에서는 아직까지 종이문서의 활용을 완전히 대체하는 것은 어렵다고 판단하고 있습니다.

프랑스의 경우는 언제까지 종이 기록의 완전한 소멸을 예상하시고 있는지, 그리고 현재 종이문서가 꼭 필요한지 여부에 대한 검사장님의 생각을 듣고자 합니다.

이상의 몇 가지 질문으로 토론을 대체하고자 합니다.

감사합니다.

토론문

[제2주제] 첨단기술의 발전과 범죄, 그리고 형사사법의 대응에 관한 토론문

이원상
(조선대학교 법학과 교수)

1. 들어가며

오늘 이렇게 뜻깊은 자리에서 토론자로 참여할 수 있는 기회를 주셔서 감사드립니다. 발제자께서는 2025년 10대 미래 기술(MIT 테크놀로지 리뷰) 등 기술의 급격한 발전이 범죄의 지능화·고도화를 초래하고 있음을 전제하며, 크게 인공지능(AI), 자율주행차, 그리고 로봇 및 웨어러블 기기를 이용한 범죄와 그에 따른 수사 대응의 쟁점을 논의하고 있습니다. 해당 분야에 대해 오랫동안 연구를 수행해 오시며 우수한 연구성과를 만들어 오신 발제자의 발제 내용에 대해 전적으로 공감합니다. 토론자가 오랫동안 생각해 왔지만 결론을 내리지 못한 사안들에 대해 발제자께서는 혜안을 제시해 주셨습니다. 발제자의 발제문을 보면서 궁금했던 내용들을 중심으로 토론을 진행하겠습니다.

2. 첨단기술을 이용한 범죄의 유형

윤지영 박사님의 발제문을 살펴보면, 기술의 발전이 범죄의 ‘지능화’와 ‘고도화’를 초래하고 있음을 알 수 있습니다. 특히 생성형 인공지능(Generative AI), 자율주행차, 로봇 기술 등이 악용되면서 기존에 없던 새로운 유형의 범죄들이 등장하고 있는데, 이를 형사법적 관점에서 크게 세 가지 범주로 나누어 볼 수 있을 것입니다.

(1) 인공지능(AI)을 이용한 지능형 범죄

가장 두드러지는 변화는 범죄의 접근성이 높아졌다는 점입니다. 생성형 AI의 등장으로

전문 지식 없이도 손쉽게 범죄 도구를 만들 수 있게 되었습니다. 먼저 딥페이크(Deepfake) 악용 사례가 있습니다. AI 기술로 가짜 영상이나 음성을 만들어 허위 사실을 유포하거나 음란물을 제작하는 범죄가 심각합니다. 특히 충격적인 사실은 딥페이크 성착취물 피해자의 국적 분석 결과, 한국이 53%로 전 세계에서 가장 높은 비율을 차지하고 있다는 점입니다. 미국(20%), 일본(10%) 등과 비교해도 압도적으로 높은 수치로, 우리 사회가 직면한 심각한 문제입니다. 또한 고도화된 피싱(Phishing)이 있습니다. 과거와 달리 소셜 미디어 정보를 AI로 분석하여 개인 맞춤형 피싱을 시도하거나, 유명인의 목소리와 얼굴을 사칭하여 투자를 유도하는 등 수법이 교묘해지고 있습니다. 챗GPT 같은 도구가 범죄 조직의 '사기 필수템'으로 전락하여 사투리를 흉내 내는 등 스캠(Scam)의 효율적 도구로 쓰이기도 합니다. 그리고 시장 교란 및 사이버 공격이 강력해지고 있습니다. 금융이나 주식 시장의 패턴을 분석해 시장을 교란하거나, 시스템 취약점을 분석해 사이버 공격의 성공률을 높이는 데 AI가 활용됩니다. 이는 결과에 있어 단순히 범죄에 국한되지 않고, 국가적 재앙을 일으킬 수도 있습니다.

(2) 자율주행차를 이용한 범죄(테러 및 강력범죄)

자율주행차는 운전자가 없다는 특성으로 인해 새로운 형태의 범죄 도구가 될 수 있습니다. 먼저, 무인 테러 수단이 될 수 있습니다. 2016년 니스 테러와 같이 대형 차량을 이용한 돌진 테러가 자율주행차로 행해질 경우, 범인은 현장에 없으므로 자신의 신체적 위험 없이 대규모 살상을 저지를 수 있게 됩니다. 범죄가 발각될 가능성 또한 낮아집니다. 또한 해킹을 통한 살인・납치가 가능해 집니다. 차량 시스템을 해킹하여 주행 경로를 조작함으로써 탑승자를 납치하거나 감금할 수 있으며, 고의적인 사고를 유발해 살인하는 도구로 악용될 수 있습니다. 그리고 범죄 공간으로 활용될 수 있습니다. 자율주행차 내부가 성매매 등 은밀한 범죄를 저지르는 공간으로 활용될 가능성도 제기됩니다. 자율주행자동차의 크기가 더욱 커지게 되면 보다 다양한 범죄 도구로 활용될 수 있으며, 추적도 쉽지 않을 수 있습니다.

(3) 로봇 및 드론을 이용한 물리적 범죄

이제는 로봇 기술을 이용하는 물리력을 행사하는 범죄로 이어질 수 있습니다. 가장 우려되는 것으로 킬러 로봇(Killer Robots)이 있습니다. 2016년 댈러스 경찰이 범인 제압을

위해 폭탄 로봇을 투입하거나, 러시아가 IS 조직원 사살을 위해 로봇을 이용한 사례가 있습니다. 이는 국가 기관의 대응 사례이지만, 역으로 범죄 조직이 이러한 로봇을 이용해 출입문을 폭파하고 침입하거나 요인을 암살할 위험이 존재합니다. 또한 드론의 위협이 우리집 문 앞에 서 있게 되었습니다. AI 드론이 사람을 노리거나 위협하는 방식으로 범죄에 악용될 수 있는 시대가 도래했습니다. 무엇보다 드론은 누구나 쉽게 구입해서 개조하고, 사용할 수 있으므로 그 위험성이 매우 높습니다. 이처럼 첨단기술은 우리 삶을 편리하게 하지만, 동시에 범죄자들에게는 추적을 피하고 범죄 효율을 높이는 강력한 무기가 되고 있습니다. 따라서 이러한 기술적 진보에 대응할 수 있도록 현행 법적 간극을 어떻게 메울지 더욱더 고민하고 노력해야 할 것입니다.

3. 질의 내용

이처럼 발제자께서 지적하신 첨단기술의 발전 속도에 비해 법제도의 정비는 다소 지체되고 있는 것이 현실입니다. 이와 같은 상황을 고려해서 해당 분야에서 많은 선도적 연구를 수행해 오신 발제자께 형사법적 관점에서 궁금한 점 세 가지를 여쭙고자 합니다.

첫째로, 자율주행차 압수・수색 시 ‘영장 제시’의 실질화 방안과 제3자 프라이버시 보호에 관해서입니다. 발제자께서는 무인 자율주행차에 승객이 탑승해 있을 경우, 수사기관이 차량에 대한 영장만으로 승객의 소지품까지 수색하는 것에 대한 비판적 견해를 소개하셨습니다. 또한 관리자에게 전자영장을 제시하는 등의 대안을 언급하셨습니다. 현행 형사소송법상 영장 제시는 피처분자의 방어권 보장을 위한 핵심 절차입니다. 만약 해킹이나 테러 혐의로 긴급하게 주행 중인 자율주행차를 정지시키고 압수・수색해야 할 때, 차량 소유주(피의자)가 현장에 없고 선의의 탑승객만 존재한다면, ’관리자(제조사/운영사)에 대한 통지’만으로 헌법상 적법절차 원칙을 충족했다고 볼 수 있는지, 그리고 이때 탑승객의 별도 소지품(프라이버시)을 보호하기 위한 구체적인 입법적 가이드라인은 무엇이라고 생각하시는지 발제자의 고견을 여쭙고자 합니다.

둘째로, 위치정보 프라이버시와 제3자 법리(Third-Party Doctrine)의 적용 한계에 관해서입니다. 발제자께서는 미국 연방대법원의 “Carpenter v. United States” 판결을 인용하며, 자율주행차 위치정보 취득의 영장주의 적용 여부를 문제 제기하셨습니다. 기존 자동차는 프라이버시 기대치가 낮아 영장 없는 수색이 광범위하게 인정되었으나, 자율주행차는 단순한 이동 수단을 넘어 거주 공간의 성격도 띠게 됩니다. 자율주행차가 생성하는 운행 정보(위치, 차내 음성, 영상 등)는 제조사나 통신사 서버에 저장됩니다. 수사기

관이 이를 압수할 때, 기존의 ‘제3자 법리(정보주체가 제3자에게 제공한 정보는 프라이버시 보호가 약화된다)’를 적용하여 영장 발부 요건이 완화되어야 할 것인지, 아니면 자율주행차 내부를 ’주거’에 준하는 공강이나 자율주행자동차를 ‘휴대폰’에 준해 프라이버시 요건을 더욱 엄격하게 해석하여 더욱 엄격한 영장주의를 적용해야 한다고 보시는지 고견을 여쭙고자 합니다.

셋째로, 치안용 스마트 안경 도입 시 법률유보 원칙과 오남용 통제 방안에 관한 질문입니다. 발제자께서는 스마트 안경을 이용한 얼굴 인식 기능이 국민적 반발에 부딪힐 것으로 전망하셨습니다. 하지만 현재 중국 등 일부 국가에서는 이미 활용 중입니다. 만약 강력범죄 예방 등 지극히 제한적인 목적하에 스마트 안경(안면 인식 기술)을 도입한다면, 현행 규정상 단순히 경찰관 직무집행법의 개정만으로 가능한 것인지, 아니면 생체 정보 처리와 관련된 별도의 특별법 제정 등이 필요하다고 보시는지 궁금합니다. 아울러 현장 경찰관이 스마트 안경을 운용할 경우 예상되는 오남용은 어떤 것이 있으며, 그를 방지하기 위한 사전적·사후적 통제 장치로 어떤 것을 고려해 볼 수 있을 것인지 고견을 부탁드립니다.

4. 나가며

사실 오늘 발제문에 대해서는 전적으로 동의하고, 발제자와 같은 생각을 하고 있습니다. 질의 내용은 토론자가 오랫동안 고민하는 숙제에 관해 해당 연구 분야의 전문가이신 발제자의 고견을 구한 것입니다. 오늘 발제자께서는 “우리 문명은 앞으로도 더욱 찬란한 미래를 맞이할 것이다!“라는 기술적 낙관론 이면에 존재하는 어두운 측면들과 법적 쟁점들에 관해 예리하게 파헤쳐 주셨습니다. 첨단기술은 계속해서 발전해 가지만, 범죄의 도구로 악용되면 인류에 큰 재앙이 될 수 있습니다. 첨단기술이 인간에게 이롭고, 범죄의 도구가 되지 않도록, 법학계가 선제적으로 규범적 기준을 마련해야 한다는 점에 깊이 공감하며 토론을 마치고자 합니다. 이렇게 뜻깊은 자리에서 토론을 할 수 있는 기회를 주셔서 다시 한번 깊은 감사를 드립니다.

토론문

[제2주제] 첨단기술의 발전과 범죄, 그리고 형사사법의 대응에 관한 토론문

류부곤
(경찰대 법학과 교수)

1. 토론을 시작하며

산이 높으면 계곡이 깊어지는 것과 같이 기술발전의 고도화는 범죄대응에 있어서도 이전과는 차원이 다른 심화된 문제를 야기합니다. 오늘의 발표는 형사사법이 대응해야 할 기술발전이 매우 넓은 영역에서 수준높은 형상으로 다가오고 있음을 잘 보여주고 있습니다. 하지만 각국의 형사사법기관이 열심히 노력하고 있음에도 기술의 발전속도를 따라가기는 쉽지 않다는 것도 확인할 수 있습니다. 저는 윤박사님이 소개해 주신 몇 가지 첨단기술의 영역에 대해 범죄와의 관련성과 대책을 중심으로 코멘트하는 것으로 토론에 갈음하도록 하겠습니다.

2. 인공지능의 범죄활용 현상에 대해

ChatGPT와 같은 생성형 인공지능이 출현하였을 때 이미 많은 학자들이 인공지능이 범죄의 도구로 악용될 것이라는 것을 예측하였습니다. 딥페이크 음란영상물과 가짜뉴스는 인공지능이 범죄의 직접적 도구가 되는 예이고, 이에 대해서는 이미 다양한 형사사법적 대처방안이 시도되고 있습니다. 하지만 다량의 정보를 빠르게 처리하고 가공하여 새로운 창작물을 만들어낼 수 있는 생성형 인공지능의 무한한 능력은 범죄에 대한 활용에 있어서도 새롭고 창의적인 영역을 만들어낼 수 있음을 의미하기에, 범죄에 적절히 대응해야 하는 형사사법시스템에게 인공지능의 범용화는 크나큰 문제가 아닐 수 없습니다. 윤박사님이 소개해주신 동남아 범죄조직의 사례는 이러한 문제를 대표적으로 보여줍니다.

이에 대해서는 우선 두 가지 정도의 접근법을 생각해 볼 수 있습니다. 하나는 인공지

능이 범죄에 악용되지 못하도록 원천적으로 차단할 수 있는 감시·감독기능을 강화하는 것입니다. 물론 이것은 형사사법기관이 독자적으로 수행할 수는 없고, 법제화 등을 통해 인공지능 서비스 사업자에게 의무를 부여하는 방식입니다. 이용자가 사기와 관련한 일정한 이용패턴[1]을 보이는 경우 적극적으로 계정이용을 중단하고 수사기관에 통보하는 등의 감독과 보고의무를 부여하되, 이 과정에서 이용자의 프라이버시 침해나 공익목적 연구자의 연구활동이 제약되지 않도록 감시방식과 인증시스템 등을 꼼꼼히 검토할 필요가 있습니다. 둘째는 인공지능이 위험한 도구임을 인정하고 이를 악용한 이용자에게 보다 가중된 형사책임을 부여하는 것입니다. 칼과 같은 흉기를 이용한 경우 맨손으로 폭행하는 것보다 가중하여 처벌하는 것과 같이 인공지능을 이용한 사기 등의 범죄행위를 가중하여 처벌하는 법규정을 신설하는 것입니다. 자동차가 상당한 위험성을 가진 수단임에도 그 편의성으로 인해 현대사회에 보편화되었듯 인공지능도 지금 시대를 살아가는 사람들에게 삶의 모든 영역에서 필수도구가 될 것이라는 점은 분명합니다. 자동차의 운전자에게 광범위한 주의의무가 부여되듯 인공지능 이용자에게도 폭넓은 주의의무가 부여되고 잘못된 결과에 대해서는 중한 책임이 인정되어야 합니다.

3. 자율주행차에 대한 압수수색 관련

중국이나 미국의 일부 도시에서 상용화된 로보택시와 같은 무인자율주행자동차의 경우 범죄와의 관련성이 인정되어 경찰에 의한 압수·수색의 대상이 된 경우, 한국의 경우는 이에 대한 영장주의 관련 규정이 존재하지 않는 상황입니다. 종래 한국의 대법원은 압수·수색의 대상이 된 주거 등의 장소에 관리자가 존재하지 않는 경우, 영장을 제시하지 않고도 집행할 수 있다고 판결한 바 있고, 이후 형사소송법에 관련 규정이 추가되었지만(제118조) 이는 여전히 예외적 상황에 대해 물리적 영장을 전제로 하는 규정입니다. 기술의 발전은 사람이 아닌 인공지능이나 로봇에 의해 관리되고 이용되는 여러 공간을 만들어내고, 영장의 제시라는 법적 활동도 얼마든지 전자적 방식으로 적법하게 이루어질 수 있습니다. 발표자료에서는 2021년 도입된 전자영장이 이용될 수 있을 것이라고 하였지만, 아직 현실은 그렇지 못합니다. 2021년의 형사절차전자문서법은 2025년 10월부터 전자영장이 사용될 수 있음을 규정하였지만, '전송'의 방식으로 영장을 제시할 수 있는 경우는 "전자적 형태로 송신 또는 수신될 수 있는 정보를 대상으로 하는 경우"[2]로 제한

1) 예를 들면 매우 반복적이고 자동화된 방식의 사기문구 생성 반복, 특정 국가(동남아 강제노역 지역 등)에서 대량 계정 동시 접속, 다수의 피해자에게 동일 패턴의 메시지를 보내는 행동 등.
2) 대법원 규칙 제36조 제2항.

되어 있습니다(이는 금융거래정보나 통신정보 등을 대상으로 압수수색을 하고자 하는 경우입니다).[3] 따라서 한국의 경우 형사소송법이나 관련 법령에 영장을 전자적 방식 등으로 원격으로 제시·전송할 수 있는 근거규정이 부재한 상황입니다. 영장의 제시는 형사처분을 받는 대상자의 기본권을 보호하기 위한 것이므로 기술의 발전을 고려하여 기술적·전자적 방식으로도 적절히 이루어질 수 있도록 법령의 정비·보완이 필요합니다.

4. 뇌-컴퓨터 인터페이스(BCI)와 행위개념 재정립 필요성

뇌와 컴퓨터를 연결하여 사람의 신경활동 만으로 일정한 행위가 가능하도록 하는 BCI 기술의 발전은 형법상 행위개념과 형사책임의 부여에 있어서 근본적인 변화를 요구할 수 있습니다. 사람의 신경신호를 감지하고 분석하여 이를 전자적 신호로 전환하는 기술은 현재 신체의 근육세포와 기계를 연결하는 방식(신경-근육 인터페이스, Neuromuscular Interface)과 뇌의 신경신호 발생조직과 컴퓨터를 직접 연결하는 방식(뉴럴링크)으로 나누어 볼 수 있는데, 행위개념과 직접적으로 관련있는 것은 후자의 방식일 것입니다. 예를 들어 사람이 생각을 하고, 이에 연결된 컴퓨터가 정보통신망에서 일정한 정보를 생성하는 행위를 하였는데, 이것이 범죄에 해당하는 경우(해킹 등의 사이버공격, 혐오나 음란표현, 정보유출 등) 이는 '생각이 곧 행위(Thinking is Doing)'인 상황이 되므로 순수한 신경활동으로서 내면의 생각과 이를 외부에 표현하는 행위를 구별하는 전통적 행위개념으로는 법적 규율이 곤란한 상황입니다. 이에 대해서는 이러한 기술이 상용화될 경우, 인공지능의 경우와 마찬가지로, 그로 인한 위험부담도 이용자에게 부과하는 것을 원칙으로 하는 접근법이 가능하겠지만, 생각에 대한 통제는 '생각만큼' 쉽지 않다는 점에서 이를 섣불리 인간의 행위로 포섭하기는 쉽지 않을 것 같습니다. 이 문제는 기술의 취지를 고려하여 이원적으로 접근할 필요가 있습니다. 즉 신체기능의 결손으로 인해 이 기술이 반드시 필요한 경우와 그렇지 않은 경우로 나누어 볼 필요가 있다는 것입니다. 사지마비 환자와 같이 인간의 기본적 생존을 위해 이 기술을 이용하는 경우, BCI 기술로 구현된 '범죄'는 이용자의 행위로 간주되어서는 안되고 기술적 '사고'로 간주되어야 합니다. 하지만 인간의 능력을 증강하거나 특정한 목적을 위한 기술적 수단인 경우에는 생각의 발현을 행위로 의제하는 법적 조치가 필요합니다. 이는 범죄에 악용된 도구의 한 종류로 보아야 하고, BCI 장치를 이용하는 선택행위 자체를 법적 주의의무가 부여된 행위로 보아야 합니다.

---

3) 이마저도 시행시기는 다시 늦춰져서 2026년 9월 28일부터 시행예정이다.

5. 토론을 마치며

이상에서 몇 가지의 첨단기술 영역과 범죄와의 관련성을 형사법학자의 시각으로 검토하여 보았습니다. 이러한 문제들에서 공통되는 것은 범죄에 대응하기 위한 형사사법적 대응책을 마련하기 위해서 기술에 대한 높은 이해가 필요하다는 점입니다. 하지만 형사사법시스템을 연구하고 운영하는 대부분의 사람들은 자연과학이나 공학에 대한 학문적 배경이 거의 없는 '사회과학자'입니다. 이 상황에서 형사사법시스템을 연구하고 정책을 마련하는 사회과학자나 법학자는 어떠한 선택을 해야 할까요? 당연히 첨단기술을 연구하고 개발하는 과학자들과 협업을 활발히 해야 합니다. 그러나 그것으로는 부족해보입니다. 과학자들과 법학자들의 언어와 사고는 차이가 많이 납니다. 과학자들이 아무리 잘 설명해도 법학자는 법학자의 사고방식으로 그것을 받아들이고 이를 법규나 정책에 반영합니다. 과학자가 이렇게 만들어진 법규나 정책을 검증할 수 있으면 좋겠지만 그것이 가능할지는 의문입니다. 그래서 법학자도 결국 공부해서 첨단기술을 이해해야 합니다. 오늘 많은 형사법학자들이 이곳에 모인 이유 중에는 내가 공부해야 할 것이 무엇인지 숙제를 받아가는 의미도 분명히 있다고 생각합니다. 감사합니다.

Panel Discussion

AI and the Future of Criminal Justice: Issues and Challenges

Moderator: Han, Sang Hoon

Professor, Law School of Yonsei University

Advisor, Korean Criminal Law Association

사회: 한상훈

연세대 법학전문대학원 교수

한국형사법학회 고문

Discussion Papers

Kim, Sung-Ryong

Professor, Kyungpook National University Law School
President, Korean Association of Criminal Procedure Law

Choi, Ho-Jin

Professor, Department of Law, Dankook University
President, Korean Association of Comparative Criminal Law

Kim, Han-Kyun

Senior Research Fellow, Korean Institute of Criminology and Justice
President, Korean Society of Criminology

토론문

첨단기술과 형사법 국제세미나 토론문

김성룡
(경북대 법학전문대학원 교수, 한국형사소송법학회 회장)

안녕하십니까? 한국형사소송법학회장 경북대학교 김성룡 교수입니다.
오늘 한, 중, 독, 프 4개국의 형사법 전문가분들의 발제와 토론 잘 들었습니다. 현재 각국의 인공지능 관련 연구의 동향, 쟁점들, 그리고 미래의 과제들에 대한 논의 현황에 대한 최신 정보를 제공해 주신 점에 대해서도 깊이 감사드립니다.

첨단기술과 형사사법의 미래, AI 시대의 범죄예방과 데이터 보안, 첨단기술의 발전과 범죄, 그리고 형사법의 대응이라는 주제에서 드러나고 있듯이, 오늘 발제와 토론의 관심 방향은 주로 발전하는 현대의 기술이 어떤 범죄를 가능하게 할 것이며, 이에 대해 형사사법은 형사실체법은 물론이고 수사와 증거, 공판과 형집행 등 형사절차법에 걸친 대응 방법을 어떻게 첨단기술 발전에 적응시킬 것인가가 핵심인 것으로 보입니다.

앞선 발제자분들과 토론자분들의 의견에 대해 세부적인 질문과 설명을 요구하는 것은 시간상의 제약으로 힘들 것이라는 생각에 필자가 늘 혼자만의 난제로 생각하고 있는 몇 가지 고민에 대해 외국의 학자분들은 어떤 생각을 하시는지, 해당 국가에서는 이에 대한 구별되는 논의나 관심이 주목받고 있는지에 대해 답변을 청하고자 합니다.

첫째는 첨단과학기술의 발전을 그대로 두고 볼 것인가하는 것입니다. 예를 들어 핵무기의 개발과 사용은, 그것이 설령 강대국의 횡포의 대상이라고 하더라도, 범세계적으로 이를 통제하는 것에 다수의 의견이 모여 있는 것으로 알고 있습니다. 사람을 살해하는 조폭 로봇, 전쟁에 투입되어 죽지 않는 불멸의 전투 로봇은 이미 생산되어 실전에 투입되었거나 조만간 투입될 것이라는 진단이 있듯이, 현재 우리의 관심의 대상이 되는 기술들

은 어떤 기준으로, 어떤 방법으로, 어떻게 개발과 사용이 제한되어야 하는 것은 아닌지, 이를 위해 국제적 공조가 시급히 이루어져야 하는 것은 아닌지에 대해 어떻게 생각하시는지요? 현대형 프랑켄슈타인의 출현을 그대로 두는 것이 인류를 위한 이익에 부합하는 것인지요?

둘째는 인공지능과 로봇의 발전이 가져올 문제는 헤아릴 수 없이 많겠지만, 현재 ANI 수준에서 AGI 수준으로, 이어서 ASI 수준을 발달한 인공지능이 장착된 인간의 피부와 구별하기 어려운 로봇이 등장할 때, 우리가 생각하는 법익보호 중심의 현행 형사법체계가 그대로 유지될 수 있을 것인지, 어디에서 무엇부터 새롭게 형사법체계를 만들어야 할지에 대해 어떻게 생각하시는지요? 예를 들어 AGI 또는 ASI 수준의 인공지능이 장착된 로봇을 강간하는 일은 불가능한 것인지, 가능하다고 한다면 이제 인간이 아닌 로봇도 범행의 객체인 피해자로 등장하는 것인지? 단지 그 로봇의 제작자, 소유·점유자, 법적·사실적 권리자에 대한 범죄로 보는 것으로 족한 것인지? 만약 ASI가 장착된 로봇을 다른 로봇들이 보는 앞에서 잔인하게 부숴버리거나 파괴하는 행위(아래 동물보호법은 동물에 대한 그런 행위를 징역 2년 이하 또는 2천만원 이하의 벌금에 처하는 규정입니다)는 형사법적 처벌의 대상이 되는 것인지? 그 근거는 무엇이며 보호법익은 무엇인지? 등의 문제에 대한 구체적인 논의나 정책적 논의가 진행되고 있는지 궁금합니다.

대한민국 동물보호법 제10조(동물학대 등의 금지) ① 누구든지 동물을 죽이거나 죽음에 이르게 하는 다음 각 호의 행위를 하여서는 아니 된다. 1. 목을 매다는 등의 잔인한 방법으로 죽음에 이르게 하는 행위 2. 노상 등 공개된 장소에서 죽이거나 같은 종류의 다른 동물이 보는 앞에서 죽음에 이르게 하는 행위 제97조 벌칙 ② 다음 각 호의 어느 하나에 해당하는 자는 2년 이하의 징역 또는 2천만원 이하의 벌금에 처한다.

생각해 보지도 못한 문제, 질문의 방향을 바꾸어야 할 문제, 여전히 고전적인 물음이 정확한 받을 제공하는 문제 등등 아직도 전혀 분석되지 못한 다양한 문제들을 가능한

빨리 음미하고 방치해서는 안 될 위험을 사전에 차단하는 노력도 오늘 여기 모이신 우리 형사법학자들의 과제의 하나가 아닌가 하는 주제넘을 생각을 전하면서 토론을 마칩니다. 감사합니다. 끝

토론문

인공지능과 형사법의 미래: 책임의 주체에서 위험사회의 통제 원리로

최호진
(한국비교형사법학회장, 단국대학교 법학과 교수)

안녕하십니까. 저는 단국대학교 최호진 교수입니다. 오늘 이 뜻깊은 국제학술대회에서 독일, 중국, 프랑스, 그리고 한국을 대표하는 학자들의 고견을 듣게 되어 큰 영광입니다. 각국의 법적, 문화적 배경 위에서 인공지능(AI)이라는 공통된 화두를 어떻게 형사법적 관점에서 풀어내고 있는지 확인할 수 있는 매우 유익한 시간이었습니다.

저 역시 이 자리에 계신 여러 발표자분들과 마찬가지로, 급변하는 기술 환경 속에서 형사법이 지켜야 할 가치와 나아가야 할 방향에 대해 오랫동안 고민해 왔습니다. 저는 그동안 주로 “인공지능은 범죄를 저지를 수 있는가?”, “인공지능에게 법인격을 부여할 수 있으며, 나아가 형법상 범죄주체가 될 수 있는가?”라는 연구질문을 가지고 몇 편의 논문을 발표한 적이 있습니다. 저의 주된 관심사는 “과연 스스로 판단하고 행동하는 것처럼 보이는 자율적 AI에게 인간과 같은 형사책임을 물을 수 있는가?”라는 근본적인 질문이었습니다.

저의 연구 결론은 “아직은 아니다”입니다. 강한 인공지능(Strong AI)조차도 -아직 강한 인공지능이 없다고 생각하지만-, 형법이 전제하는 '자유의지'에 입각한 윤리적 비난가능성, 타행위가능성을 전제로 한 비난가능성을 핵심개념으로 하는 '규범적 책임'을 부담할 수 있는 주체로 보기는 어렵다는 것입니다. AI에게 섣불리 독자적인 형사책임 주체성을 인정하는 것은 근대 형법의 대원칙인 책임주의를 훼손할 위험이 있습니다. 따라서 저는 AI 자체를 처벌의 대상으로 삼기보다는, AI가 야기한 결과에 대하여 그 배후에 있는 설계자, 제조자, 혹은 이용자의 과실 책임과 주의의무의 범위와 그 분배에 대하여 논의를 하는 것이 현재의 형법 해석론과 책임 원칙에 부합한다는 입장을 견지하고 있습니다.

오늘 세 분 석학의 발표는 저의 이러한 고민의 지평을 넓혀주었습니다. 각국의 발표 내용 중 저의 연구 방향과 맞닿아 있으면서도 새로운 시사점을 준 부분에 대해 깊은 공감을 표하고 싶습니다.

독일의 에리히 막스 교수님의 발표는 형법의 근본 가치를 다시금 상기시켜 주었습니다. 교수님께서는 AI 시대에도 인간의 존엄과 책임주의라는 형법의 대원칙은 결코 흔들려서는 안 된다고 역설하셨습니다. 특히 국가의 효율적인 AI 활용이 자칫 시민의 자유를 과도하게 침해하는 '감시 사회'로 이어질 가능성에 대한 준엄한 경고는, 우리가 기술 도입 과정에서 반드시 견지해야 할 인권적 기준점을 제시해주었습니다.

중국의 천징춘 교수님께서 지적하신 'AI의 법적 지위에 대한 신중한 접근'에 전적으로 동의합니다. AI의 독자적 행위 능력을 인정하면서도 이를 성급히 법인격 부여로 연결하는 것을 경계하고, 기존의 제조물 책임이나 사용자 책임 법리를 통해 해결책을 모색하려는 단계적 접근은 형법의 보충성 원칙에 비추어 매우 타당하다고 생각합니다.

프랑스의 에릭 마테 검사장님의 발표는 형사사법 실무의 최전선에서 느끼는 AI의 양면성을 생생하게 보여주었습니다. 수사 도구로서의 유용성 이면에 있는 '블랙박스' 문제와 기본권 침해 우려에 대한 지적, 특히 사법적 판단의 최종 권한은 여전히 인간에게 있어야 한다는 강조점은 기술만능주의에 대한 중요한 메시지입니다.

저는 앞으로 인공지능과 형사법 영역에서 우리가 주목해야 할 방향성을 제시하기에 앞서, 몇 가지 오래된 전설을 들려드리고자 합니다.

먼저 16세기 프라하의 '골렘' 전설입니다. 당시 랍비 유다 뢰브는 박해받는 유대인 공동체를 보호하기 위한 선한 의도로 진흙에 생명을 불어넣어 거대한 인조인간 골렘을 만들었습니다. 처음에는 사람들을 돕던 골렘은 점차 통제를 벗어나 폭주하며 도시를 파괴하는 재앙이 되었고, 결국 랍비는 눈물을 머금고 자신이 만든 피조물을 다시 흙으로 되돌려야만 했습니다.

괴테의 시 '마법상의 제자'(Der Zauberlehrling)와 월드 디즈니 영화 '환타지아'에 유명한 이 이야기도 있습니다. 위대한 마법사의 제자가 스승이 자리를 비운 사이, 자신이 통제할 수 없는 강력한 마법을 부려 빗자루에게 물을 길어오게 시킵니다. 빗자루는 명령대로 물을 계속 길어오지만, 제자는 멈추는 주문을 모릅니다. 집안은 물바다가 되고, 제자가 도끼로 빗자루를 내리치자 빗자루가 둘로 나뉘어 두 배의 속도로 물을 길어 나릅니다. 제자는 자신이 불러낸 힘에 의해 익사할 위기에 처합니다. 마법사는 즉시 주문을

풀고 물난리를 멈춥니다.

한국에서도 어깨너머로 어설프게 도술이나 주술을 배운 선비나 젊은이가 있었습니다. 그는 책에서 본 대로 부적을 태우고 주문을 외워 강력한 '귀신 장군'이나 '신장(神將)'을 소환했습니다. 하지만 그는 소환 주문만 알았지, 그들을 다루거나 돌려보내는 방법은 전혀 몰랐습니다. 나타난 귀신장군은 통제에서 벗어나 주변의 죄 없는 사람들까지 닥치는 대로 해쳤다는 한국의 민담도 있습니다.

이 이야기들은 "자신의 능력을 과신하고 아무런 안전장치 없이 감당할 수 없는 위험한 힘이나 존재를 불러내는 행위가 얼마나 어리석고 파멸적인 결과, 즉 자신뿐만 아니라 주변까지 고통받는 결과를 초래하는지"를 경고하는 이야기들입니다.

이제 우리는 논의의 중심을 'AI가 범죄의 주체가 될 수 있는가'라는 실체법적 책임의 문제를 넘어, 'AI라는 새로운 위험을 어떻게 관리할 것인가'라는 위험 통제의 문제로 확장시켜야 합니다. 무엇보다 우리는 인공지능의 본질이 결국 인간의 기술 발전과 편익 증진을 위한 강력하고 유용한 도구적 성격에 있다는 점을 잊지 말아야 합니다. AI는 자율적 의지를 가지고 스스로 목적을 설정하고 행위하는 주체가 아니라, 인간의 목적을 달성하기 위한 수단입니다.

형법학적 관점에서 볼 때, 현대 사회의 인공지능은 막대한 효용과 함께 내재적 위험성을 동시에 지닌, 이른바 '허용된 위험원(危險源)'입니다. 자동차나 원자력처럼 과학기술의 발전과 사회적 유용성 때문에 그 존재를 허용하지만, 언제든 심각한 법익 침해를 야기할 수 있는 존재라는 의미입니다. 우리는 지금 이 거대한 위험원과 공존해야 하는 고도화된 '위험사회'를 살아가고 있습니다.

안전사회에 대한 지향과 위험 통제의 관점은 이미 국제적인 입법 흐름에서도 구체화되고 있습니다. 최근 유럽연합(EU)이 마련한 '인공지능법(AI Act)'이 대표적인 예입니다. 이 법은 AI 시스템이 국민의 안전과 기본권에 초래할 수 있는 위험의 수준을 '수용 불가능한 위험', '고위험', '제한적 위험' 등으로 분류하고 그에 따라 차등적인 규제를 가하는, 철저한 '위험 기반 접근방식(Risk-Based Approach)'을 주요 내용으로 하고 있습니다. 따라서 향후 형사법적 논의의 핵심은 이 '허용된 위험'의 '법적 허용 범위', 제조사, 프로그램 개발자 등 인공지능 관련자들의 주의의무 범위를 어디까지로 설정할 것인가에 모아져야 합니다. 이는 단순한 기술적 기준을 넘어선 다층적인 논의가 필요합니다.

이제 위험사회의 형사법이 감당해야 할 구체적인 과제는 명확합니다.

첫째, 법학적 관점에서 '허용된 위험의 규범적 한계'를 명확히 획득해야 합니다. 이는 단순히 기술의 효용과 위험을 저울질하는 이익 형량의 문제를 넘어섭니다. 제 연구의 핵심 결론이기도 하듯이, AI라는 도구가 야기한 결과에 대해 '배후에 있는 인간—설계자, 제조자, 이용자—의 주의의무 범위와 책임 소재를 어떻게 합리적으로 분배할 것인가'를 정교하게 법제화하는 작업이 선행되어야 합니다. 어디까지가 사회적으로 용인 가능한 '적절한 위험'이고, 어디서부터가 법적 책임을 물어야 할 '과실'인지 그 경계선을 긋는 것이야말로 미래 형사법의 가장 시급한 책무입니다.

둘째, 사회적 관점에서는 기술에 대한 신뢰의 위기를 극복해야 합니다. 알고리즘의 본질적인 불투명성(블랙박스)은 대중에게 막연한 불안과 공포를 심어줍니다. 형사사법 절차에 AI가 도입된다면, 이 '깜깜이 판결'을 국민이 과연 신뢰할 수 있을까요? 따라서 알고리즘의 투명성을 확보하고 설명 책임을 강화하여 기술에 대한 최소한의 사회적 신뢰 기반을 마련해야 합니다. 이러한 신뢰 없이는 어떠한 법적 통제도 사상누각에 불과할 것입니다.

결국 미래의 형사법은 인공지능이라는 강력한 도구가 우리가 감당할 수 없는 '통제 불가능한 거대한 힘'으로 돌변하여 인류를 위협하지 않도록, '인간의 책임을 명확히 하는 법률적 통제'와 '투명성을 담보로 한 사회적 합의'라는 두 축을 바탕으로 '안전한 규범의 울타리'를 공고히 하는 시대적 과업에 주력해야 할 것입니다.

오늘 이 자리가 기술의 발전이 인간의 존엄과 가치를 훼손하지 않고, 오히려 정의로운 형사사법 실현에 기여할 수 있도록 지혜를 모으는 소중한 계기가 되기를 희망합니다. 경청해 주셔서 감사합니다.

2025.11.24.

최호진

토론문

AI-형사정책의 전망들에 대한 회고

김한균
(한국형사정책학회장, 한국형사·법무정책연구원 선임연구위원)

1. 인공지능기술의 형사정책적 활용 내지 형사사법시스템에서 인공지능기술 도입의 필요성과 수요, 그 긍정적 효과에 대한 기대와 부정적 영향에 대한 우려 자체도 기술발전과 사회변화에 따라 변화한다. 더구나 그 변화가 빠르고 광범하기 때문에, 가까운 과거의 전망을 되돌아보면 현재의 상황을 조망하고, 멀지 않은 장래의 변화를 전망하는데 도움이 될 것이다.

2. 2000년대 초반 형사정책 분야에서 인공지능 활용 전망 내지 기대는 주로 컴퓨팅 파워와 빅데이터 분석을 활용해 범죄방지의 효과를 높이는 데 집중되었다.

2.1. 예측형 치안(predictive policing)은 가장 관심이 집중된 분야로서, AI 알고리듬을 활용해 과거 범죄 데이터(시간, 장소, 유형)를 분석하여 범죄 발생 가능성이 가장 높은 특정 지역과 시간대인 “hot sopt”을 예측하는 것이 목표였다. 이러한 예측을 활용해 경찰 순찰 경로와 자원 배치를 최적화하고, 범죄를 사전에 억제하며 대응 시간을 단축하기 위해 고위험 지역에 경찰관을 선제적으로 배치하는 것이었다. NYPD CompStat(Compare Statistics)와 같은 초기 프로그램은 2010년 후반에 등장한 PredPol (현재 Geolitica)과 같은 더 진보된 AI 기반 시스템의 토대를 마련했다.
2.2. 컴퓨터 비전과 신경망기술기반 차량번호판 자동인식(Automatic License Plate Readers)기술은 자동화된 공공장소 감시를 통하여 법집행기관의 범죄예측과 예방에 활용될 수 있다.
2.3. 개인의 범죄 위험을 객관적으로 평가하기 위해 데이터와 알고리즘을 활용하는 범죄위험평가도구는 법관에게 데이터 기반 평가를 제공함으로써 양형판단, 보석결정의 객관

성을 확보하는데 활용될 수 있다. 또한 형사사법체계 행정업무를 간소화하고 방대한 법률 문서를 처리함으로써 수사와 기소, 그리고 사법행정의 효율성을 높이는 데 기여할 것으로 기대되었다.

3. 그러나 첨단과학기술을 활용한 형사정책에 대해 효율성뿐만 아니라 공정성 제고효과까지 낙관적 전망과 기대도 있었지만, 이후 현재까지 더욱 깊어질 문제점에 대한 우려와 비판도 제기되었다.

3.1. AI 모델이 과거 범죄와 범죄자 데이터를 학습하는 동시에, 종래 경찰활동 또한 이미 특정 지역과 인구 집단에 불균형적으로 집중되어 있었기 때문에 AI 도구는 기존의 체계적 편향성을 오히려 고착시키고 증폭할 위험까지 있다.
3.2. AI 알고리즘 작동방식의 블랙박스 측면으로 인해 그 예측 결과를 검토하거나 오류에 대한 책임을 묻기 어렵게 된다.
3.3. 자동화된 감시의 확장과 개인 데이터 수집 능력 증대에 따라 국가의 일상 감시 가능성에 대해 시민의 프라이버시와 자유 침해 위험성에 대한 문제가 제기된다.

4. 21세기 첫 4분기를 지나온 현 시점에서 이러한 문제의식은 더 강력한 규제와 투명성 확보 필요성에 대한 법적, 사회적, 문화적 요구로 이어지고 있다.

4.1. 데이터 편향에 대한 초기 우려는 다양한 AI 애플리케이션, 특히 위험 평가 도구(보석, 형량, 가석방 결정) 및 예측 치안 시스템에서 이어지고 있다. COMPAS(Correctional Offender Management Profiling for Alternative Sanctions) 위험 평가 도구는 유사한 범죄 기록을 가진 백인 피고인에 비해 흑인 피고인을 불균형적으로 고위험군으로 분류한다는 문제가 드러났다. 해법의 초점은 편향성 식별에서 공정성 지표 개발 및 적용(인구통계학적 균등성, 확률 균등화 등)과 데이터 선택에 있어서의 새로운 윤리적 프레임워크 구축으로 전환되었다. 즉 인공지능기술이 기존 불평등을 증폭시키기보다 완화하는 데 사용되도록 하기 위한 대책이다.
4.2. 정교한 머신러닝 모델의 사용으로 투명성과 유해한 결과에 대한 AI 시스템의 책임 문제도 더욱 복잡해지고 있다. 얼굴 인식이나 복잡한 위험산정(risk scoring) AI 시스템의 낮은 설명 가능성(XAI)은 적법절차 원칙과 충돌하게 된다.
4.3. 예측형 치안 모델의 정확도, 범죄감소 효과에 대한 평가 결과 현장에서의 사용이

축소되는 결과에 이르기도 한다. AI 시스템 기반 결정이 잘못된 체포, 부당한 판결 또는 권리 침해 결과를 초래할 경우, 현 법체계로는 소프트웨어 개발사, 법집행기관, 운영자 사이에서 책임성은 여전히 문제다.

4.4. 인공지능기반 감시기술 중에서 공공장소에서의 얼굴인식기술, 디지털 증거에 대한 신뢰라는 개념 자체를 위협하는 딥페이크 기술, 소셜 미디어, 공공 기록, 상업 데이터를 활용해 개인 프로파일링에 활용하는 기술은 시민의 권리뿐만 아니라 형사사법제도의 공정성에 더욱 위협이 되고 있다.

4.5. 2000년대 초반 잠재적 위험성에 대한 우려는 이제 구체적 피해에 대한 대응의 문제로 심화되고 있다. 형사사법제도 뿐만 아니라 사회제도 전반에서 인공지능기술 사용이 민주주의 가치와 헌법적 권리체계에 부합하도록 보장하기 위한 입법적 대응이 강조되고 있다. 2024년 세계 최초의 유럽연합 인공지능기본법에 이어 2026년 1월 시행 예정인 한국의 인공지능 발전 및 신뢰성 기반 구축에 관한 기본법(인공지능기본법)이 그 예다. 두 법제 모두 위험 기반 접근법을 공유하지만, 근본적으로 다른 규제 철학을 반영한다. 즉 EU는 위험 완화와 기본권 보호를 우선시하는 반면, 한국은 국가 경쟁력 확보를 위해 규제와 적극적인 산업 진흥 간의 균형을 명시적으로 추구한다.

5. 형사정책의 역사에서 새로운 범죄방지기법과 기술의 도입을 둘러싸고 언제나 기대와 낙관, 우려와 비판이 있어왔다. 이제까지 형사정책학이 담당한 중요한 역할이기도 했다. 형사사법체계에 있어서 인공지능기술과 그에 기반한 기법, 도구들은 종래 기술과 도구와 다른 차원의 문제를 제기하게 될 것인가?

과거의 전망을 회고해 보고, 현재와 비교해 보면서, 다시 장래를 전망해 보건대 기술이 문제의 프레임을 바꾸기 보다는 언제나 문제였던 과제의 프레임 안에서 더 나은 방법을 찾고 부정적 영향을 줄여나가는 과제는 변함이 없을 것이다. 다시 말해서 형사정책의 변함없는 과제는 국가형벌권력으로부터 시민의 자유를 보장하고, 범죄와 범죄피해로부터 시민의 안전을 보호하는 것이다. 어떠한 범죄방지제도와 정책수단이 그러했듯 그 자체가 안전과 자유 양 측면에 긍정적 효과와 부정적 영향을 함께 가져온다. 긍정적 측면을 증진하고 부정적 측면을 피하기 위해서는 언제나 인간의 권리, 시민의 기본권을 기준으로 정책판단을 해야 한다는 사실에는 변함이 없다.