2022년
한국형사정책학회
전북대학교 법학연구소
한국형사·법무정책연구원
춘계공동학술대회

AI와 형사정책

| 일시 | 2022년 3월 25일(금)
| 장소 | 전북대학교 법학전문대학원 501호/810호

한국형사정책학회 Korean Association Of Criminology
전북대학교 법학연구소 LAW RESEARCH INSTITUTE OF JEONBUK NATIONAL UNIVERSITY
KICJ 한국형사·법무정책연구원 Korean Institute of Criminology and Justice

한국형사정책학회
Korean Association Of Criminology

인사말

반갑습니다. 한국형사정책학회가 전북대학교 법학연구소 및 한국형사법무정책연구원과 공동하여 개최한 춘계학술대회에 오신 여러분을 환영합니다. 저는 한국형사정책학회 학회장을 맡고 있는 문성도입니다.

저희 학회 운영 방향은 1985년 한국형사정책학회 창립취지문 끝자락에 잘 나타나 있습니다.

> "범죄연구는 다른 어떤 분야보다도 정책적 관심과 참여가 높은 분야이며, 따라서 우리는 기존 입법 및 정책에 대한 분석, 비판 및 대안의 모색을 아울러 추구하고자 한다. 이러한 추구에 있어 과학적 탐구가 결여된 형사정책은 맹목적이며, 때로는 편의주의적인 정책이 되기 쉬운 반면, 인권 내지 사회정의의 이념에 이끌리지 않는 형사정책은 국민을 통제대상으로만 여기는 태도에 다음 아니다. 다시 말하면 '과학'과 '인권'이 형사정책의 양대 지주가 되어야 함을 우리는 확신한다."

한국형사정책학회는 1985년 설립 이후 다양화·심각화되는 범죄현상에 대응하여 그 실태와 원인을 분석하고, 범죄에 대한 기존 입법 및 정책을 분석하고 비판하는 데 그치지 않고 구체적이고 실질적인 형사정책 방안을 제시함으로써 형사정책 분야를 양적·질적으로 발전시키는데 앞장서 왔습니다. 한국형사정책학회의 학술지인 「형사정책」은 법학관련 총 157개의 등재학술지중 인용지수 7위이며 형사법 관련 등재학술지 중에서는 단연 으뜸입니다. 이것은 그만큼 형사정책 분야에 대한 연구 수요가 있고 한국형사정책학회가 그에 대한 역할을 하고 있다는 증거일 것입니다.

2022년 올 한 해도 저희 학회는 과학적인 실증연구를 기반으로 인간의 권리가 보장되고 사회 정의가 강물처럼 흐르는 세상을 만들기 위해 형사정책 분야에서 계속 노력하고자 합니다. 형사정책 분야는 그 어느 분야보다 학제간 및 학문상호간의 교류와 협력 그리고 공동작업이 요구됩니다. 저희 학회는 형사정책의 전문가들이 함께 모여 자유로이 의견을 교환하고 신진학자들이 기꺼이 자신들의 연구결과를 펼쳐놓고 다 함께 논의하는 장을 계속 마련하고자 합니다.

오늘의 대주제는 "AI와 형사정책"입니다. 제1주제 "고위험 인공지능기술의 윤리적 규율과 형사법적 통제"에 대해 본 학회 조직이사이신 한국형사·법무정책연구원 김한균 박사님께서 발표해주시고 본 학회 법제이사이신 조선대 이원상 교수님께서 토론해 주십니다.

제2주제 "인공지능을 이용한 재범위험성평가의 의의와 한계"에 대해 고려대 장진환 박사

님께서 발표해주시고 본 학회 이사이신 한림대 박노섭 교수님께서 토론해 주십니다.

제3주제 “AI를 활용한 구속기준의 재정립”에 대해 동의과학대 김병수 교수님께서 발표해 주시고 본 학회 국제이사이신 전남대 최병천 교수님께서 토론해 주십니다.

제4주제 “인공지능 윤리가이드와 형사정책적 시사점”에 대해 “EU 인공지능법과 인공지능 윤리가이드를 중심으로” 계명대 김희정 교수님께서 발표해주시고 본 학회 홍보이사이신 전북대 조기영 교수님께서 토론해 주십니다.

제5주제 “성폭력 피해자 면담 조사용 챗봇 개발과 형사정책적 함의”에 대해 동국대 조은경 교수님과 양기주 교수님께서 공동 발표해주시고 본 학회 재무이사이신 제주대 안성조 교수님께서 토론해 주십니다.

오늘 이 학술회의는 인공지능에 대한 과학적 탐구를 통하여 인간을 위한 형사정책적 제안을 제시하는 장이 될 것입니다. 코로나사태로 인하여 이 학술행사장에 참석하지 못하는 분들을 위하여 화상회의를 동시에 진행하고 있습니다. 이 학술행사에 적극 동참하고 협력하여 주셔서 오늘 이 행사가 좋은 결실을 맺게 되길 기대합니다. 학술회의 발제와 토론·사회를 맡아주시고 또한 이 자리에 함께 하신 여러분께 감사드립니다.

끝으로 한국형사정책학의 발전에 평소 관심을 가져주시고 특히 춘계학술회의 개최를 공동으로 개최해주신 한국형사법무정책연구원 하태훈 원장님과 김민영 박사님 등 관계자 여러분, 전북대 법학연구소 소장 김태명 교수님과 안근욱 선생님 등 관계자 여러분께 진심으로 감사를 드립니다. 춘계학술회의를 기획하고 진행해주신 오병두 연구이사님과 이진수 총무간사님께도 감사를 드립니다. 이 자리에 함께 하신 여러분, 세상을 뒤덮은 감염병에 맞서 승리하여 건강하고 행복한 삶을 누리시길 기원합니다. 감사합니다.

2022년 3월 25일

한국형사정책학회 학회장

문성도

KICJ 한국형사·법무정책연구원
Korean Institute of Criminology and Justice

축 사

안녕하십니까?

한국형사·법무정책연구원 원장 하태훈입니다.

따스한 봄날 전북대학교에서 2022년 한국형사정책학회 춘계 공동학술대회를 개최하게 된 것을 진심으로 축하합니다.
강의와 연구, 업무로 바쁘실 텐데도 직접 혹은 온라인으로 참석하신 학계와 실무계 회원 여러분을 진심으로 환영합니다. 오늘 뜻깊은 학술행사를 공동으로 주최할 기회를 주신 문성도 한국형사정책학회 회장님과 김태명 전북대학교 법학연구소 소장님께 감사드립니다. 학술행사장에 직접 참석하지 못하고 이렇게 화상으로 인사드리게 된 점 매우 송구합니다.

팬데믹 사태가 지속되면서 우리 일상의 삶과 사회적 환경에 많은 변화가 일어나고 있습니다. 이제 다시 일상을 회복하고 사회통합의 제도와 정책을 전향적으로 준비해 나가야 할 때입니다. 우리 형사법과 형사정책 연구자들은 사회적 거리두기 상황 속에서도 온라인으로 함께 모여 한국 사회의 현안과 미래 과제를 공부하는 노력을 계속해 왔습니다.

한국형사·법무정책연구원 또한 2017년 “법과학 적용 형사사법 선진화 방안(III): 인공지능 기술”, “암호화폐(Cryptocurrency) 관련 범죄 및 형사정책 연구”를 시작으로, 2018년 “4차산업혁명 시대 형사사법적 대응 방안(I): 자율주행 자동차와 드론”, 2019년 “제4차 산업혁명 시대의 형사사법적 대응 및 발전방안(Ⅱ): 사물인터넷(IoT)과 블록체인”, 2021년“산업별 인공지능 융합 촉진을 위한 법제 대응 방안”등 4차 산업혁명 시대를 대비한 선제적·선도적 연구를 다수 수행해 왔습니다.

오늘 학술대회 주제인 “인공지능과 형사정책”에 관한 토론의 장은 팬데믹 상황에서 더욱 급진전하는 이른바 디지털 대전환의 핵심 내용인 인공지능과 관련하여 다양한 현안과 과제를 함께 논의해 보는 뜻깊은 자리가 될 것입니다.

특히 지난 2020년 정부의 “인공지능 법·제도·규제 정비 로드맵”에 따르면, 인공지능 확산에 따른 부작용은 알고리즘·데이터 편향성, 취약계층의 인공지능 이용격차 증가, 플랫폼 노동자의 고용 불안정, 개인정보 침해, 거대 빅테크 기업의 시장지배력 강화 등으로 예상됩니다. 이에 따라 중장기적·포괄적 인공지능 법·제도·규제 정비의 필요성이 제기되고 있습니다. 형사법과 형사정책적 연구가 충실하게 뒷받침되어야 할 이유입니다.

바로 오늘 논의될 주제인 고위험 인공지능기술의 통제와 윤리적 가이드라인, 인공지능을 활용한 재범 위험성 평가, 구속기준, 성폭력 피해자 면담 조사 문제는 인공지능 형사정책의 방향과 연구과제를 제시해 주리라 기대됩니다. 또한 이번 학술행사를 계기로 첨단과학 기술의 사회적 영향에 관한 형사정책 연구에 있어서 우리 연구원과 학회와의 창조적인 협력관계를 더욱 발전시켜 나갈 수 있기를 기대합니다.

끝으로 발표와 토론, 사회를 맡아주신 여러분께 감사의 인사를 드립니다.

고맙습니다.

2022년 3월 25일

한국형사·법무정책연구원

원장 하태훈

2022년 춘계공동학술대회 일정

※ Zoom 접속은 다음 링크를 참고하시면 됩니다.
- 기조발제 및 제1세션 : 회의 ID: 974 1179 1206 / 암호: 1237
- 제2세션 : 회의 ID: 978 5967 1814 / 암호: 1237

<table>
<tr><td>13:00-13:10</td><td colspan="2">접수 및 장내 정리</td></tr>
<tr><td>13:10-13:20</td><td colspan="2">개회사</td></tr>
<tr><td>13:20-13:30</td><td colspan="2">축사 : 김태명(전북대학교 법학연구소장)</td></tr>
<tr><td>13:30-13:40</td><td colspan="2">축사 : 하태훈(한국형사·법무정책연구원장)</td></tr>
<tr><td rowspan="3">13:40-14:40</td><td colspan="2">기조발제</td></tr>
<tr><td colspan="2">사회: 김태명 (전북대)</td></tr>
<tr><td colspan="2">고위험 인공지능기술의 윤리적 규율과 형사법적 통제
발표: 김한균 (한국형사·법무정책연구원)
토론: 이원상 (조선대)</td></tr>
<tr><td>14:40-14:50</td><td colspan="2">휴 식</td></tr>
<tr><td rowspan="4">14:50-15:50</td><td>제1세션</td><td>제2세션</td></tr>
<tr><td>사회: 김태명 (전북대)</td><td>사회: 오병두 (홍익대)</td></tr>
<tr><td>인공지능을 이용한
재범위험성평가의 의의와 한계</td><td>AI를 활용한 구속기준의 재정립</td></tr>
<tr><td>발표: 장진환 (고려대)</td><td>발표: 김병수 (동의과학대)</td></tr>
<tr><td></td><td>토론: 박노섭 (한림대)</td><td>토론: 최병천 (전남대)</td></tr>
<tr><td>15:50-16:00</td><td colspan="2">휴 식</td></tr>
<tr><td rowspan="3">16:00-17:00</td><td>인공지능 윤리가이드와 형사정책적 시사점 : EU 인공지능법과 인공지능 윤리가이드를 중심으로</td><td>성폭력 피해자 면담 조사용
챗봇 개발과 형사정책적 함의</td></tr>
<tr><td>발표: 김희정 (계명대)</td><td>발표: 조은경 (동국대),
양기주 (동국대)</td></tr>
<tr><td>토론: 조기영 (전북대)</td><td>토론: 안성조 (제주대)</td></tr>
<tr><td>17:00-17:10</td><td colspan="2">폐회사</td></tr>
<tr><td>17:10-20:00</td><td colspan="2">석 식</td></tr>
</table>

기조발제

< 기조발제 - 발 표 문 >

고위험 인공지능기술의 윤리적 규율과 형사법적 통제

김한균 (한국형사·법무정책연구원 선임연구위원)

I. 인공지능의 위험과 그 의미

> "생각하는 기계라는 새로운 위험은 훨씬 가까이 와 있습니다. ... 인간의 특징 중에 어떤 것은 결코 기계로 모방할 수 없으리라며 일말의 위안을 찾으려는 경우도 많습니다. 이를테면 어떤 기계도 훌륭한 글을 쓸 수 없다거나 성적 매력에 이끌릴 수 .... 없다 할 겁니다. 하지만 저는 어떤 위안도 드릴 수 없습니다. 그런 한계를 둘 수 없을 것이기 때문입니다. 하지만 인체를 닮은, 분명히 인간적이지만 지적이지는 않는 기계를 만드는 .... 시도는 헛수고로 그칠 것이고, 그 결과는 造花처럼 불쾌할 것입니다."[1]

70년전 오늘날 인공지능 구현 방법을 최초로 제시했던 튜링에 따르면, 첫째, 인공지능은 새로운 위험을 가져온다, 예상보다 빨리. 1951년의 튜링은 다음 천년기 이내[2]일 것이라 보았지만, 21세기가 이미 '인공지능의 시대(Age of AI)'다. 인공지능은 일개 산업이나 제품, 또는 과학분과를 넘어, 그 학습능력과 진화적 특성에 기해 제조, 물류, 교통, 국방, 법집행, 정치, 예술, 광고, 연구, 교육에 이르기까지 인간 삶의 다양한 측면에서 급격하고 광범위한 전환을 초래하며, 인간의 근대적 정체성과 경험체계까지 변화시킬 것이기 때문이다.[3]

둘째, 실로 인간지능 특징의 기계적 구현에는 한계를 둘 수 없다. 이미 GPT-3(Generative Pre-trained Transformer-3)는 사람처럼 글을 작성하고 대화를 하며[4], 성기능 로봇(sex robot)이 유통되기 시작했다.[5]

셋째, 다만 인간적 지능을 구현하거나 인간보다 더 뛰어나게 생각하는 기계보다는 단순히 인간기능의 일부를 구현하고 강화한 기계부터 문제가 될 것이다. 이미 등장한 소피아[6]나 제미노이드(Geminoid) HI-4[7]

---

1) Alan Turing, Can Digital Computers Think? (1951년 5월 BBC 라디오 강연) 앨런 튜링, 지능에 관하여, 노승영 역, 2019, 133면.
2) Alan Turing, Can Digital Computers Think? 133면.
3) Henry Kissinger·Eric Schmidt·Daniel Huttenlocher, The Age of AI and Our Human Futues, 2021, 8면.
4) 개발자 OpenAI 에 따르면 GPT-3는 뉴스기사를 작성할 수 있고 대화응답도 가능하다. 하지만 학습데이터에 따라서는 혐오발언도 할 수 있다. (Matthew Hutson, Robo-writers: the rise and risks of language-generating AI, Nature 2021.3.3)
5) "State of the Sexbot Market: The World's Best Sex Robot and AI Sex Doll Companies"(https://futureofsex.net/robots/state-of-the-sexbot-market-the-worlds-best-sex-robot-and-ai-love-doll-companies/ 2022년 3월 1일 최종검색)
6) 2015년 개발된 휴머노이드 로봇으로 인터넷 챗봇 기술이 적용되어 인간처럼 자연스러운 대화기능을 목표로 한다. 카메라로 사람의 얼굴표정을 읽고, 손으로 제스처를 취하면서 비언어적 대화까지 가능하다. (David Hambling·Liron Gilenberg, *WE: Robot,* 2018, 백승민 역, 우리는 로봇이다, 2019, 210면.)
7) Erico Guizzo, *The Man Who Made a Copy of Himself,* IEEE Spectrum 47(4) 2010; 2013년 일본의 공학자 히로시 이시구로가 자신을 정밀하게 복제한 로봇이다. 인간의 얼굴표정과 움직임을 표현하는 원격조정 로봇으로,

의 경우 조화처럼 불쾌한 반응을 유발할 수 있고, 나아가 인간적 형태를 입고 성기능이나 살상기능만을 수행하는 기계라면 불안의 대상이 될 수 있다.

1. 어떻게 위험을 이해할 것인가

물론 '한계를 둘 수 없는' 인공지능에 대해 과학기술이 넘을 수 없는 한계를 말하거나, 먼 장래의 문제라 미룬다 한들 위안이 되지는 못한다. 한편으로는 인공지능의 실존적 위협이 어떤 구체적 위험으로 결과될지 예견하기도 쉽지 않다. 인공지능기술 위험에 대한 사회적 평가는 과장과 경시 사이에서 아직 합리적이고 합의가능한 자리를 자리잡지 못했다. 인공지능기술 전문가들도 기술전도사(technology evangelist)와 예언자(prophet)으로 갈려 장밋빛 미래와 잿빛 종말에 대해 각각 설파하는 가운데 대중매체는 이를 극적으로 전파하고 있는 상황이다.[8)]

스티븐 호킹과 빌 게이츠는 "완전한 인공지능의 개발은 인류의 종말을 가져올 수 있다."[9)] "인공지능은 우리 실존에 대한 최대 위협이기 때문에 국가적·국제적 차원에서 일정한 규제감독이 필요하다."[10)]고 경고한다. 반면 노동과 에너지와 자원의 희소성을 넘어선 유토피아를 전망하거나,[11)] 오히려 인공지능기술과 인간이 결합된 증강인간(Augmented Human)이나 초인간(Trans-human)을 추구하기도 한다.[12)]

이처럼 상반된 두 입장의 공통점은 과대평가라는 점이다. 비관론은 새로운 기술을 맞아 오래된 두려움을 소환하고 있다.[13)] 인공지능 위험에 대한 불안 정서는 현대과학기술의 광범한 혜택에도 불구하고 안전관리체계가 부실하거나 오작동할 경우 초래될 위험은 파멸적이어서 국가와 법체계가 위험을 통제하고 시민을 보호해 주리라는 믿음이 흔들리게 되는 상황[14)]의 연장인 측면도 있다. 범용인공지능(AGI)이 현실적으로 가능하지 않거나 예상가능한 장래에 실현되지 않는 한[15)] 자의식과 고유의 욕망을 지닌 인공지능체가 인류를 위협하는 일은 있을 수 없다. 반면 낙관론은 그리고 인공지능기술 발전이 사회적 위기이자 기회가 될 수 있다는 말은 위기와 기회 분별이 불투명하다는 뜻이기도 하다는 사실, 인공지능의 학습과 작동방식이 인간이해 수준을 넘어섰다는 사실[16)], 그리고 인간과 기계의 하이브리드화(hybridization)로 지능과 수명을 월등히 높인다해도 기술권력과 불평등 문제는 여전하다는 사실[17)]을 간과한다. 그리고 인공지능기

인간과 상호작용속에서 uncanny valley 효과 (지나치게 인간과 비슷한 형체에 대해 급격히 호감도가 떨어지고 불안감이나 혐오감이 야기되는 효과)를 탐구하는데 활용된다. (David Hambling·Liron Gilenberg, WE: Robot, 217-218면)

8) "현실세계에서 인공지능이 실제 할 수 있는 일은 과대평가되고 있다. 언론보도가 대단치 않은 진전도 대전환을 가져올 것인양 과장하기 때문이다. 마치 14세기에 사는 사람들이 교통사고를 걱정하는 꼴이다. 14세기라면 공중위생에 대해 걱정하는 편이 훨씬 유용하지 않을까?" (Gary Marcus·Ernest Davis, Rebooting AI, 2019, 이영래 역, 2029 기계가 멈추는 날, 2021, 36-37면)

9) "Stephen Hawking warns AI could end mankind" (BBC News 2014년 12월 2일자)

10) "With AI, We are summoning the demon"(Washington Post, 2014년 10월 24일자)

11) Aaron Bastani, Fully Automated Luxury Communism, 2019, 김민수·윤종은 역, 완전히 자동화된 화려한 공산주의, 2020, 76면.

12) Thomas Powers & Jean-Gabriel Ganascia, "The Ethics of the Ethics of AI", in M.Dubber et al, eds. T*he Oxford Handbook of AI,* 2020, 36면.

13) Thomas Powers & Jean-Gabriel Ganascia, The Ethics of the Ethics of AI, 34면.

14) 김한균, 후기현대사회의 위험관리를 위한 형법 및 형사정책 연구(I), 현대과학기술사회 위험관리 형법 및 형사정책의 체계와 원리, 한국형사정책연구원, 2012, 39면.

15) Michael L. Littman et al, Gathering Strength, Gathering Storms: AI100: 2021 Study Panel Report, 2021, 30면.

16) Davide Castelvecchi, "Can we open the black box of AI?" NATURE 538, 2016, 20-23면

술을 위험하게 활용하거나 범죄도구로 악용하려는 욕망을 지닌 인간은 언제나처럼 존재한다.
사실 생각하는 기계나 범행하는 기계(Criminal AI)[18]가 인류의 장래를 위협하는 문제라기보다는, 인간의 신체적, 정신적 기능을 모사하여 인간의 기능을 대체하거나 강화하는 기계를 우리 일상의 삶 속에 수용하는 과제야말로 실제 다룰 의미가 있는 위험한 인공지능의 문제다. 현실적으로 기계지능(machine intelligence)은 인간과 협업하거나 보조지원하는 능력 개선방향으로 개발되고 있다.[19] 이는 인공지능의 위험이 프랑켄슈타인이나 스카이넷와 같은 형태가 아니라, 인간과 상호작용관계를 바탕으로 일상적 삶속에 밀접한 영향을 미칠 수 있게 될 자동화의사결정시스템(automated decision-making system)이나 특정기능성 로봇의 형태로, 그리고 개인의 정체성, 지향과 취향 정보가 인공지능의 도구가 될 수 있는 형태로 제기될 것임을 말해준다.

2. 어떤 위험을 통제할 것인가

인공지능 기술의 잠재적 위협이나 현실 위험성에 대한 잘못된 평가는 예측오류와 규제실패의 문제로 이어진다는 점에서도 문제다. 예컨대 19세기말 도시 교통수단인 말 5만마리 때문에 런던이 50년뒤 2.7미터 높이의 배설물 아래 파묻힐 것이라는 언론의 경고와 학계의 대책논의가 있었다. 1894년 이른바 말똥위기(horse Maure Crisis)는 불과 30년뒤 마차가 자동차로 대체되면서 잘못된 우려였음이 드러났다. 한때 기술적 난제로 보였던 문제 상당수는 생각보다 빨리 왜 문제 삼았었는지 조차 의아할 지경이 될 수 있다.[20] 마차에서 내연기관 자동차로의 전환이 인공지능기술에 기반한 자율주행수단으로의 전환과 단순비교 되기는 어렵지만, 위험예측과 대응은 언제나 오류가능성을 열어두고 신중해야 한다는 점은 변함없다.
특히 과장 인식된 위험은 더 많고 더 강한 통제 필요성에 대한 사회적 압력을 높인다는 점에서 과소평가해서는 안된다. 불충분한 정보에 근거한 부적절하게 강화된 규제는 인류에게 혜택을 줄 기술적 가능성마저 차단하는 부당한 결과에 이를 수 있기 때문이다.[21] 법제도를 포함해 위험을 방지하기 위해 투입할 사회적 자원 역시 한정되어 있다는 점도 고려해야 한다.
그래서 인공지능기술 위험방지와 윤리적·법적 대응은 문제지점을 적절히 파악하고, 그에 대응한 규제를 정당하고도 효과적으로 배치하기 위한 숙고의 과정이 될 것이다. 인공지능 기술에 내재된 또는 인공지능 기술로 초래될 위험의 실체여부, 실체적 위험의 경우 그 정도에 따른 분별에 기해 통제 방안을 도출된다. 윤리적 규율과 법적 통제 대상인 인공지능기술 위험과 관련하여 현행 지능정보화기본법이 규정한 '국민의 생명 또는 신체안전 등에 밀접한 지능정보기술'[22]을 정부의 '인공지능 법·제도·규제 정비 로드맵'은 고위험 인공지능기술 위험예방 규정이라 이해하고, 디지털 포용구현과 복지 분야에 인공지능 도입을 위한 법적 기반 확충 필요성을 제시한다.[23] 2021년 '알고리즘 및 인공지능에 관한 법률안'[24]은 고위험인공

17) Thomas Powers & Jean-Gabriel Ganascia, The Ethics of the Ethics of AI, 36면
18) Eric Horvitz, One Hundred Year Study on Artificial Intelligence: Reflections and Framing, 2014, 3면.
19) Michael L. Littman et al, Gathering Strength, Gathering Storms, 24면.
20) Aaron Bastani, Fully Automated Luxury Communism, 106-109면.
21) Peter Stone et al, Artificial Intelliegence and Life in 2030: AI 100: Report of the 2015-2016 Study Panel, 2016, 10면.
22) 지능정보화기본법은 고위험분야(국민의 생명·신체 등에 밀접한 지능정보기술) 사업자가 지능정보기술을 개발·관리·활용 시 준수 의무가 있는기술기준 고시에 관한 근거 규정(제21조)을 두었으나, 현재 인공지능기술 발전 수준과 인공지능 산업 관련 시장의 성숙도 등을 고려하여, 기술기준 고시 제정은 유보 중이다. (관계부처합동, 인공지능 법·제도·규제 정비 로드맵, 2020, 33면.)

지능을 '국민의 생명, 신체의 안전 및 기본권의 보호에 중대한 영향을 미치는 인공지능' (동법안 제2조 제3호)이라 정의한다.
이는 유럽연합의 위험(평가)기반 규제개별화 (risk-based, sector-specific approach)방식과 방향을 같이 한다.[25] 즉 안면인식 프로그램(facial recognition software)과 같은 고위험 인공지능기술의 개별 활용에 한정하고 특정하여 선제적으로 안전, 공정성, 정보보호 요건을 부과하고 그 부합여부를 사전평가한다는 것이다.[26] 고위험 인공지능활용(high-risk AI applications)이란 해당 분야에서 중한 위험이 예상되는 (significant risks can be expected) 경우다. 이어서 2021년 유럽인공지능법안(legislative proposal for an Artificial Intelligence Act)은 특정인공지능기술 금지와 고위험인공지능 시스템 규제가 주요 목적이다.(제1조(a)(b))
시민의 생명과 신체의 안전과 인권 보장에 대한 상당한 위험을 기준으로 합의·제시된 위험방지의 규칙과 법제는 인공지능기술에 대한 불신과 불안에 대한 해법이 될 수 있다.[27] 다만 고위험 인공지능기술 통제는 형법적 제재도 필수적으로 상정해야 하기 때문에, 그 제재의 과잉이나 부실, 공백 또한 문제에 대응한 제재가 오히려 문제의 일부가 될 위험도 높다는 딜레마를 대면해야 한다.

3. 어떻게 위험을 통제할 것인가

기술사학자 크란츠버그의 기술법칙(Laws of technology)에 따르면,[28] 제1원칙은 기술은 선하지도 악하지도 않되, 중립적이지도 아니하다는 것이므로, 제2원칙에 따르면 개입과 규제는 필요성의 문제다. 다만 제4원칙은 기술이 공적 현안의 중심요소라 할지라도, 기술 정책은 기술외적 요소에 대한 판단이 선행되어야 한다. 인공지능기술위험에 대한 법정책도 마찬가지다. 인공지능기술은 사회적 수요로 인해 개발(공급)되기도 하지만, 기술의 공급이 필요(수요)를 만들기도 한다. 삶의 편의와 질을 혁신하는데 활용될 수도, 치밀한 감시와 치명적 침해의 전혀 새로운 수단으로 동원될 수도 있다. 따라서 고위험인공지능기술에 대한 통제 정책은 기술 자체뿐만 아니라 그 활용과 위험 문제에 대한 사회적 판단이 선행되어야 하고, 인공지능기술이 창출하거나 가중한 고위험에 대해서는 개입과 통제 필요성이 법정책적으로 검토될 필요가 있는 것이다.
물론 인공지능기술 위험에 대한 사회적 평가와 마찬가지로 윤리와 법, 특히 형법의 적절한 자리를 찾기는 어렵다. 인공지능 기술을 규율하는 윤리적 지침으로 충분할 것인가. 형사법적 통제까지 필요할 것인가. 필요하다면 그 대상과 범위는 어떻게 설정할 것인가. 윤리적 규율과 형사법적 통제 사이의 관계는 어떻게 설정할 것인가. 예측이 어려운 위험은 선제적으로 예방 또는 금지되거나, 오히려 금지를 유보해야 할 수 있고, 예상되는 위험에도 불구하고 통제속에 활용되어야 할 인공지능기술도 있다. 게다가 인류가 의지해 온 윤리와 법의 대상은 인간과 사물인데, 인간이기도 하고 사물이거나, 인간도 사물도 아닌 기술

---

23) 관계부처합동, 인공지능 법·제도·규제 정비 로드맵, 2020, 32면
24) 의안번호 13509호(윤영찬 의원대표발의, 2021년 11월 24일)
25) 미국의 2022년 알고리듬 공적책임법안 (Algorithmic Accountability Act), 캐나다의 2019년 자동화의사결정규칙 (Directive on Automated Decision-Making)도 위험기반 체계에 따라 고위험인공지능에 특화된 규제방식을 취하고 있다.
26) WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust,COM(2020) 65 final, 2020, 17-18면.
27) Peter Stone et al, Artificial Intelliegence and Life in 2030, 5면.
28) Melvin Kranzberg, Technology and history: Kranzberg's Laws, Technology and Culture 27(3), 1986, 545-559면.

구현개체나 시스템이 그 대상이 될 것이라면, 윤리적 규율, 정부 규제, 형법적 통제의 경계도 식별이 어려워질 수 있다.
하지만 이제 인공지능 시대에서 '인공지능'은 기술개발과 활용, 운용시스템을 모두 포함한 개념이며, 그 운용시스템의 핵심은 인공지능 윤리와 법제이기 때문에 그 활용의 무한한 혹은 예측불가한 가능성 안에서 해서는 안되는 것과 할 수 있어도 하지 않아야 하는 것에 대한 규범적 판단과 정책설계가 필요하고, 그에 따라 인공지능이 야기할 개별 위험을 방지할 윤리적 지침과, 특정 침해의 실질과 형태에 따라 형법적 제재를 포함한 법적 통제가 부과될 것이다. 인공지능기술 위험에 대한 통제는 위험평가 — 윤리적 지향가치 — 법적 규제의 형태로 설계된다.
이는 인공지능기술의 효용과 안전을 모두 보장하기 위한 설계로서, 종래 안전 설계(safety design)를 넘어 윤리적 설계 (ethical design) 및 적법 설계 (legitimate design)의 차원까지 포괄한다. 현실 인공지능 기술이 인간에 유사한 형태와 기능을 구현하면서 인간경험과 밀접한 상호작용을 한다면 단순히 기계제품의 위험방지를 위한 안전규제를 넘어 윤리적이고 합법적으로 기능하도록 설계하고, 또한 윤리적이고 합법적으로 활용되도록 통제가능해야 인공지능의 시대 역시 인간문명의 시대일 수 있다.
다만 인공지능 기술과 그에 기반한 시스템의 윤리적 설계, 적법 설계는 단순히 윤리학자와 법률가가 제시한 내용을 도덕엔진 또는 법률데이터 코드에 담아 기술적으로 탑재[29]하는 문제차원이 아니다. 과학기술 연구개발의 설계단계부터 윤리적·법적 관점에서 인공지능의 기능과 역량을 구현하여야 한다면, 무엇이 인류사회의 보편적 윤리와 인공지능 기술에 고유한 윤리인지, 어떻게 이에 부합되게 인공지능 기술을 개발·운용하고, 어떻게 이에 반하는 위험을 규정하고 규율·통제할 것인지에 대한 사회적 합의가 선행되고, 이러한 사회적 합의를 어떻게 법제도적으로 보장할 것인지가 인공지능 국가전략의 최우선과제가 되어야 한다.
인공지능 국가전략의 일부로서 형사정책은 인간의 인공지능 기술활용에 따른 위험성과 인공지능기술의 범죄적 악용에 대한 형사법적 개입을 의미할 것이다. 인공지능 형법의 과제는 고위험 인공지능기술의 위험방지, 인공지능기술의 불법적 이용으로 인한 침해제재 두 가지다. 인공지능기술이 구현된 기계, 인공지능기술 개발·운용 기업, 인공지능기술기반 공적 제도, 세 가지 영역이 이른바 인공지능 형사정책의 영역이 된다.

Ⅱ. 고위험 인공지능 대응을 위한 윤리적 가치와 법적 규제의 동향

현재 세계 최대의 전기전자공학 전문가 단체로서 관련 기술표준을 선도하는 IEEE에 따르면, 모든 형태의 인공지능 기술, 돌봄 로봇이든 자율주행운송수단이든, 또는 의료진단 소프트웨어나 알고리즘기반 챗봇이든, 공히 적용될 수 있는 보편적 윤리원칙을 정립에 있어서는 첫째, 인권 이념이 체화되어야 하고, 둘째, 인류와 자연환경의 최대 이익을 최우선적으로 고려되어야 하며, 셋째, 인공지능자율시스템이 사회시스템으로서 진화해나가면서 미치게 될 위험이나 부정적 영향을 최소화할 수 있도록 해야 한다.[30] 뿐만 아니

29) 鄭雄一, AIに「善惡の判斷」を敎える方法, 2018, 이시훈 역, 인공지능에 살해당하지 않기 위해 인공지능에 도덕엔진을 탑재하는 법 - 인공지능 로봇에게 어떻게 선악 판단을 가르칠 것인가, 2019
30) The IEEE Global Initiative for Ethical Consideration in Artificial Intelligence and Autonomous Systems, Ethically Aligned Design A Vision for Prioritizing Human Well-being with Artificial Intelligence and

라 정부가 대응해야 할 법적 과제는 자동화기계 관련 침해에 대한 기업의 책임(liability of industry for accidents involving autonomous machines) 해결이다. 침해결과에 따른 종래 법적 주체로서 자연인이나 법인 책임 법제에 대한 새로운 도전으로서 인공지능기술의 광범위한 활용전에 해결되어야 할 가장 중요한 사회적 과제이기 때문이다. 31)

현재 인공지능기술 개발과 활용이 지향해야 할 가치를 담아 책임있는 인공지능(responsible AI), 신뢰가능 인공지능(trustworthy AI), 설명가능 인공지능(explainalbe AI), 인간중심 인공지능(human-centered AI)등이 논의되고, 공적책임성(accountability), 투명성(transparency), 공정성(fairness)과 같은 윤리적 가치에 정향된(ethically aligned) 인공지능 설계와 지침, 입법논의가 다양한 형태와 방식으로 추진되고 있다.

1. 유럽연합의 인공지능 위험평가기반 규제 법체계

유럽연합은 인공지능기술이 1950년대부터 개발·활용되어왔지만 새로운 발전단계로 진입했다는 점, 그 자체 목적이 아니라 수단이라는 점을 전제하고, 크나큰 잠재적 혜택과 함께 많은 위험을 초래할 가능성에 대응해 유럽적 가치와 규칙을 존중하면서 인공지능기술을 통제하겠다는 것이다. 이에 따라 유럽이 선도적으로 인공지능기술의 개별적 위험에 상응하여 비례적이고 융통성 있는 규범을 제시함으로써 국제적 기준을 향상시키고자 한다. 그 핵심표지가 바로 '인간중심적이고, 지속가능하며, 안전하고 포용적이며, 신뢰가능한 인공지능(human-centric,sustainable, secure, inclusive and trustworthy AI)'이다.32)

따라서 유럽연합 차원 인공지능 정책의 핵심은 기술적 수월성(excellence in AI)과 신뢰가능한 기술(trustworthy AI)이다. 이는 시장과 공공부문 기능의 보장, 그리고 시민의 안전과 기본권 보호에 대한 규칙기반위에 인공지능기술의 향상이 추진되어야 한다는 의미다.33) 신뢰가능한 인공지능기술은 개발과 적용, 사용자들에게 안전하면서도 혁신적인 환경을 조성하기 위한 목표다. 이를 위해 논의되는 법적 기반은 첫째, 특정 인공지능시스템의 위험방지와 기본권 보장을 위한 법적 기본틀(legal framework for AI), 둘째, 인공지능기술 관련 책임문제 해결을 위한 규정(EU rules to address liability issues), 셋째, 관련분야 안전 규칙(Regulation) 및 지침(Directive)의 개정이다.

2021년 유럽인공지능법안은 보건, 교통, 에너지, 공공 부문 인공지능기술을 특정하여 개별고유의 규제기본틀을 새롭게 구성하며, 고위험인공지능기술을 규제하되, 정기적 평가를 통해 규제대상과 내용을 수정한다.34)

(1) 인공지능 규제기본틀

Autonomous Systems, Version 1- For Public Discussion, Institute of Electrical and Electronics Engineers, 2016, 15면.

31) IEEE, Artificial Intelligence: Calling on Policy Makers to Take a Leading Role in Setting a Long-Term AI Strategy - An IEEE European Public Policy Initiative Position Statement, 2017

32) European Commission, AI Excellence: Ensuring that AI works for people (https://digital-strategy.ec.europa.eu/en/policies/ai-people 2022년 3월 1일 최종검색)

33) "A European approach to artificial intelligence" (https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence 2022년 3월 1일 최종검색)

34) Proposal For A Regulation of The European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, 2021, para.42.

2021년 4월 유럽연합의 인공지능 규제기본틀 (Regulatory framework proposal on artificial intelligence)은 유럽연합 디지털전략 체계[35] 안에서 인공지능협업계획(Coordinated Plan on AI)과 함께 유럽차원에서 인공지능 위험에 대응하기 위한 법적 기본틀로서 제안된 것이다. 유럽을 신뢰가능한 인공지능의 글로벌허브(global hub for trustworthy AI)로 발전시키겠다는 유럽연합 정책의 핵심인 법적 기본틀과 협업계획은 시민과 기업의 안전과 기본권을 보장하고, 인공지능 투자와 혁신을 강화하는데 목적이 있다. 새 법적 체계의 방향은 신기술과 그에 기반한 다양한 새로운 애플리케이션에 대한 안전 규칙을 정립함으로써 사용자 시민의 신뢰를 높이는데 있다. 인공지능에 대한 신뢰는 바람직한 정도를 넘어 필수적인 것이기 때문에, 규제분야에서도 국제적 기준을 선도함으로써 윤리적 기술과 미래를 보장하고 혁신친화적인 법제도를 통해 시민의 안전과 기본권을 보장하는 길을 열겠다는 것이다.[36]

(2) 2021년 유럽연합 인공지능법안의 4단계 위험규정

2021년 유럽연합 인공지능법안은 인공지능에 대한 포괄적 법적 기본틀(legal framework)[37]로서는 최초다. 인공지능기술의 특정활용과 관련된 위험에 초점을 맞춘 수직적 규제(a horizontal regulation)방식을 취한다는 점,[38] 인공지능기술의 급속한 진화를 고려해 미래지향적 융통성(future-proof approach)으로써 법적 규제를 기술변화에 상응하도록 한다는 점이 특징이다. 인공지능기술의 개발과 활용 단계에서 모두 신뢰가능

35) 유럽연합은 유럽시민과 기업을 위한 디지털시대 (Digital Decade)로 규정하고 인간중심, 지속가능한 디지털 사회를 지향하는 전략을 추진하고 있다. 디지털시대에 대응하는 유럽의 방식(European way)을 규정하는 정책지표는 디지털역량을 갖춘 시민과 고급전문인력, 안전하고 지속가능한 디지털인프라, 기업의 디지털전환, 공공서비스의 디지털화(digitisation)다. 그 핵심정책영역이 바로 인공지능, 클라우드 컴퓨팅, 디지털 정체성, 데이터, 그리고 연결성(connectivity)인 것이다. 또한 유럽연합은 2022년 1월 디지털권리와 원칙 선언(declaration on digital rights and principles) 초안을 공개했다. 동 선언초안의 주요내용은 ①인권중심의 디지털전환(digital transformation), ②연대와 포용의 지지, ③온라인 공간에서 선택의 자유 보장, ④디지털공적공간에서의 참여증진, ⑤안전,안보와 개인역량의 증진, ⑥디지털 미래의 지속가능성 증진이다. (https://digital-strategy.ec.europa.eu/en/policies/europes-digital-decade#ecl-inpage-kyvdt35y 2022년 3월 1일 최종검색)

36) European Commission, “Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence” (Press release, 2021년 4월 21일자)

37) https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (2022년 3월 1일 최종검색)

38) European parliament, Artificial intelligence act, Briefing: EU Legislation in Progress, 2021, 1면. 이는 다음과 같이 (같은 문서, 5면) 위험을 단계별로 구분하여 엄격하게 단계에 상응한 인공지능기술 활용에 대해 규제하기 때무에 수직적 규제라 한다.

성이 확보되어야 하기 때문에, 기술공급자의 지속적인 질적 위험관리(quality and risk management)를 요한다.[39]

기본적으로 2021년 법안의 위험평가기반 규제방식은 인공지능기술의 위험을 ①절대적 위험(unacceptable risk), ② 고위험(high risk), ③제한적 위험(limited risk) ④ 최소위험(low or minimal risk)으로 구분한다. 그 기준은 인공지능기술의 특성으로서 불투명성, 복잡성, 데이터의존성, 인간통제를 벗어난 작동으로 인하여 사용자의 안전과 시민의 기본권에 부정적 영향을 미치는 정도다. 위험기반 규제방식은 엄격하게 각 단계 특정 위험에 대응하는 필요한 한에서 규제가 부과된다는데 특징이 있다.[40]

① 절대적 위험 인공지능 활용 (법안 제5조)

절대적 위험은 유해한 인공지능 활용(AI practice)으로 인한 침해가 명백히 금지되는 경우다. 사람의 안전, 생존, 권리에 대한 명백한 위협(clear threat to people's safety, livelihoods and rights)으로서, 금지된 위험(unacceptable risk)을 야기하는 위험을 말한다. 구체적으로는 유해한 조작적 잠재의식조종기술(manipulative subliminal techniques)사용, 장애인 등 특정 취약계층에 대한 착취, 공공기관의 사회적평판기록(social scoring) 목적, 법집행기관의 공공장소 실시간 생체정보판독 시스템(Real-time remote biometric identification systems) 관련 인공지능기술이다.

②고위험 인공지능 시스템(법안 제6조, 제7조)

고위험은 인공지능시스템이 사람의 안전과 권리에 대해 부정적 효과(adverse impact)를 초래하는 경우다. 고위험 인공지능시스템은 보건안전법규가 적용되는 제품(항공, 자동차, 의료기기, 완구)의 안전요소로 사용되는 경우와, 특정분야에 활용되는 경우로 구분된다. 고위험인공지능 시스템은 자연인에 대한 생체정보판독과 분류, 핵심인프라 관리운영(의료용 수술로봇), 교육훈련, 고용 및 노동자관리(채용서류분류 소프트웨어), 주요 공공서비스 및 부조 이용과 수혜(대출심사 신용판정), 법집행(증거능력 판단), 출입국관리 및 이민 난민(여권발급심사), 사법행정과 민주적 절차(개별사안 재판), 8개분야로 특정된다. 해당 분야는 필요에 따라 수정보완되며, 각각 새로운 고유의 규칙이 부과된다.[41]

분석해 보면, 고위험의 내용은 특히 보건복지(핵심인프라, 교육훈련,고용)과 공적절차(생체정보관리, 공공서비스이용,출입국, 법집행, 사법행정과 민주적 절차) 부문의 안전과 권리에 대한 위해가능성이다.

따라서 고위험 인공지능시스템에 대해서는 개발단계에서 엄격한 의무요건이 부과된다. 첫째, 위험평가와 감소방지시스템이 충분히 갖추어 져야 한다. 둘째, 위험유발적이며 차별적 결과를 최소화하기 위해 양질의 데이터가 시스템에 제공되어야 한다. 셋째, 결과에 대한 사후검증가능성 확보를 위해 작동기록(logging of activity)이 관리되어야 한다. 넷째, 준법(compliance) 감시평가를 위해 시스템관련 모든 정보의 상세한 기록이 관리되어야 한다. 다섯째, 사용자에게 명확하고 충분한 정보가 제공되어야 한다. 여섯째, 위험방지를 위한 적절한 수단으로서 사람의 감독이 보장되어야 한다. 일곱째, 높은 수준의 건전성(robustness), 보안과 정확성을 갖추어야 한다.

---

39) European parliament, Regulatory framework proposal on artificial intelligence (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai 2022년 3월1일 최종검색)

40) European parliament, Artificial intelligence act, 5면.

41) European parliament, Regulatory framework proposal on artificial intelligence (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai 2022년 3월1일 최종검색)

이러한 의무요건은 일차적으로 개발자의 준법감시와 행정적 규제를 통해 실현되며, 그 적절한 준수는 법적 책임과 형사제재의 면책요건이 될 수 있을 것이다. 반면 그 위반에 대해서는 행정적 제재, 나아가 중대한 위반이나 위반으로 인한 침해결과에 대해서는 형사제재가 부과될 것이다.

③제한적 위험 인공지능 시스템 (법안 제4장)
고위험 인공지능시스템에 비해 투명성 요건(transparency obligations) 부과로 한정되는 경우가 제한적 위험이다. 예를 들어 챗봇, 감정인식 시스템, 생체정보판별 시스템, 딥페이크 이미지· 영상·음성물의 상대방이 된 사람에게는 기계와의 상호작용여부를 결정할 수 있도록 상황에 대한 충분한 정보가 주어져야 한다.[42]

④허용된 최소위험 (법안 제9장)
그 밖의 인공지능시스템은 낮은 위험 수준으로 법적 의무로부터 자유롭게 허용된다. 인공지능기반 비디오게임이나 스팸 필터링 프로그램이 그 예다. 다만 2021년법안은 허용된 위험범위안에 있는 인공지능 시스템이라 하더라도 고위험인공지능시스템에 부과된 요건에 상응하는 의무를 자율적으로 이행하도록 일종의 모범준칙(codes of conduct) 필요성도 인정한다.[43]

2. 미국의 국가인공지능전략 법체계

(1) 국가인공지능전략법

2020년 국가인공지능전략법 (National Artificial Intelligence Initiative Act of 2020)은 미국 국가인공지능주도전략의 법적 기반이다. 동법상 국가인공지능주도전략(National Artificial Intelligence Initiative)[44]의 목적에는 공공 및 민간부문에서 신뢰가능한 인공지능시스템(trustworthy artificial intelligence systems)의 선도적 개발과 활용, 모든 국민을 위한 인공지능시스템의 수혜가 포함된다.[45]) 또한 2019년 국가인공지능연구개발전략계획의 8대전략[46] 중에도 인공지능의 윤리적, 법적, 사회적 의미의 이해와 대응과 인공지능시스템의 안전과 보안 확보가 포함된다.
국가기술표준연구소(National institute of standards and technology)는 인공지능 관련 협업기본틀, 기준, 지침과 관련기법, 인공지능시스템 운용에 있어서 위험방지 기본틀(risk-mitigation framework for deploying artificial intelligence systems), 신뢰가능한 인공지능시스템 증진을 위한 기술표준과 지침 개발 지원을 담당한다. 인공지능 학습데이터와 응용기술의 편향성 여부를 검사할 기술표준과 지침 개발 지원도 포함된다. 특히 2022년까지 국가기술표준연구소는 민관협력을 통해 위험관리기본틀 (Risk Management Framework)을 개발하고, 정기적으로 업데이트해야 한다. 위험관리기본틀은 신뢰가능한 인공지능시스템 개발, 인공지능시스템의 신뢰가능성 평가, 인공지능시스템의 위험방지를 위한 표준, 지침, 모범사례, 방법론

42) European parliament, Artificial intelligence act, 6-7면.
43) European parliament, Artificial intelligence act, 7면.
44) https://www.ai.gov/ 2022년 3월 1일 최종검색
45) 2020년법 Sec. 5101. National Artificial Intelligence Initiative. (A) Establishment; Purposes
46) Select Committee On Artificial Intelligence Of The National Science & Technology Council, The National Artificial Intelligence Research And Development Strategic Plan: 2019 Update, 2019, 3-4면.

과 절차를 제공하며, 활용가능한 국제표준에의 부합, 민간부문의 자율적 표준과 산업계의 모범사례 수용하는데 활용된다. 또한 신뢰가능성의 공통개념과 특징을 정립한다. 여기서 '신뢰가능성(trustworthiness)'이란 설명가능성(explainability), 투명성, 안전, 프라이버시, 보안, 견고함, 공정, 편향, 윤리, 유효성, 인증, 해석가능성 등을 포함한다.[47]

또한 국가과학재단과 국립과학공학의학 학술원(National Academies of Sciences, Engineering, and Medicine)은 컴퓨팅 기술분야 전략실행을 위한 연구 거버넌스와 관련하여 연구를 진행하고 권고안을 제시하게 된다. 이는 윤리적, 사회적, 안전, 안보의 고려가 연구설계와 검증절차에 반영됨으로써(incorporation of ethical, social, safety, and security considerations into the research design and review process) 잠재적 위해의 발생을 사전방지 할 수 있기 때문이다.[48]

특히 2020년법에 따라 설치된 법집행기관에서의 인공지능 활용문제 자문위원회(National AI Advisory Committee's Subcommittee On Law Enforcement)는 데이터편향, 데이터보안, 정보기관 및 법집행기관에서 인공지능의 도입가능성 문제와 인공지능사용에 있어서 프라이버시, 시민의 자유, 사회적 약자권리에 부합하는지에 관한 법적 기준에 대해 자문하는 역할을 담당하게 된다.[49]

(2) 인공지능 위험의 법적 규제

2019년 연방의회에 최초 상정된 알고리듬 공적책임법안 (Algorithmic Accountability Act)은 인공지능 알고리듬이 편향적, 차별적 산출물에 대한 연방정부의 규제를 내용으로 하며, 인공지능의 적법하고 윤리적인 활용에 대한 사회적 요구를 반영하고 있다. 동법안의 핵심내용은 연방무역위원회(Federal Trade Commission)에 시민 개인정보를 활용, 저장, 공유하는 개인과 기업에 대하여 영향평가(impact assessments) 및 확인된 편향 및 안전문제에 대한 합당한 해결조치에 관하여 규제를 부과하고 이를 이행하는 권한을 부여한다는데 있다. 2022년 2월 현재 2022년 법안이 재상정되어 있다.

현행법으로는 2019년 일리노이주 인공지능화상면접법(Artificial Intelligence Video Interview Act)[50] 이 있는데, 동법은 고용주로 하여금 지원자에게 화상면접에 앞서 인공지능이 면접과 채용판단여부에 활용될 수 있다는 사실을 고지하도록 의무화하였다. 또한 고용주는 각 지원자에게 인터뷰에 앞서 인공지능이 지원자 평가와 관련하여 어떻게 작동하는지 고지해야 한다. 지원자에게 인공지능기반 채용평가프로그램 사용에 대하여 사전동의도 받아야 한다. 고용주는 해당 면접영상의 불필요한 공유가 금지되며, 지원자의 요청이 있을 경우 삭제해야 한다. 2021년 인공지능화상면접법 개정안은 고용주가 지원자 대면면접여부를 판단하는데 전적으로 인공지능에 의존하는 경우 주 상업 및 경제부에 특정 인적정보를 수집보고하도록 규정하고 있다. 이는 해당정보를 분석하여 주 정부와 의회에 보고케 함으로써 인공지능 활용에 있어서 인종적 편견(racial bias)이 있었는지 점검하기 위함이다.

3. 영국의 인공지능 윤리와 안전 관리체계

47) Sec. 5301. National Institute Of Standards And Technology Activities
48) Sec. 5401. Artificial Intelligence Research And Education
49) Sec. 5104. National Artificial Intelligence Advisory Committee.
50) 820 ILCS 42

영국의 경우 인공지능 기술과 관련산업 발전상황을 고려할 때 인공지능에 특화된 규제기관이나 특별법제가 필요한지에 대해서는 합의에 이르지 못하고 있다. 영국 상원 인공지능위원회(Artificial Intelligence Committee)에 따르면 기존 법령으로 규제가 충분히 가능하다는 입장, 즉각적인 특별법제가 필요하다는 입장, 규제도입은 신중해야 하며 점진적으로 고려해야 한다는 입장으로 갈려 있다.[51] 잉글랜드 및 웨일즈 변호사협회(Law Society of England and Wales)와 기술산업협회(TechUK)는 기존 법령으로 충분하다는 입장이다. 영국 상원 인공지능위원회 역시 현 단계 인공지능에 특화된 새로운 규제법제의 필요성에 대해서는 부정적 입장이다.[52]

다만 2016년 로봇공학및인공지능 정책보고서(Robotics and artificial intelligence)는 기술 발전속도에 보조를 맞추지 못하는 관련 법적 규제의 문제점을 지적하면서, 안정적인 규제환경의 구축이 인공지능기술에 대한 공적 신뢰도를 높이는데 중요한 요소가 된다고 본다. [53] 따라서적절한 규제와 법적 기본틀의 발전은 인공지능기술 발전에도 도움이 된다. [54] 따라서 인공지능 거버넌스 구축과 관련하여 1980년대 생명윤리위원회와 같은 형태의 상설 인공지능위원회(standing Commission on Artificial Intelligence) 설립, 국립환경공해위원회(Royal Commission on Environmental Pollution)와 같은 권한을 가진 기계지능위원회(Machine Intelligence Commission) 설립방안도 제시되고 있다.[55]

(1) 2019년 인공지능 윤리 및 안전 지침

영국 인공지능청(Office for Artificial Intelligence)[56]에 따르면 인공지능기술 위험은 오용, 설계상 문제, 예측을 벗어난 위해로 구분된다. 오용은 설계 및 의도된 목적 이외의 목적에 사용되는 경우로 구분한다. 설계상 문제는 알고리듬 편향이나 안전 리스크와 관련된 기술적 문제를 철저히 고려하지 않은 경우다. 예측을 벗어난 결과적 위해는 개인과 공동체에 미치게될 부정적 가능성을 철저히 고려하지 않은 경우다.

이에 2019년 정부디지털청(Governemt Digital Service), 인공지능청과 알란튜링 인공지능연구소(Alan Turing Institute)[57]가 협업하여 종합적인 인공지능 윤리 및 안전지침을 제시하였다. [58] 인공지능 윤리지침의 우선적 목표는 다음과 같다. 첫째, 인공지능기술 허용여부는 영향을 받게 될 당사자들과 공동체 복지에 미칠 영향을 평가해야 한다. 둘째, 인공지능기술과 그 활용은 공정하고 차별적이지 않아야 한다. 개인과 사회

51) Matt Hervey and John P Cooper, *Regulation of Artificial Intelligence and Big Data in the UK*, 2019 (https://www.lexology.com/library/detail.aspx?g=4406f03c-28e3-462e-82bf-11b0bba79143 2022년 3월 1일 최종접속)

52) House of Lords, Report of Session 2017-19, 16 April 2017, HL Paper 100, para 373.

53) House of Commons, Science and Technology Committee, *Robotics and artificial intelligence*, 2017, paras.60-62

54) House of Commons, Robotics and artificial intelligence, paras.128, 135

55) House of Commons, Robotics and artificial intelligence, para.143.

56) 디지털문화미디어스포츠부(Department for Digital, Culture, Media & Sport)와 기업동력산업전략부(Department for Business, Energy & Industrial Strategy)의 연합기관으로서 인공지능청의 기본임무는 인공지능 기술이 책임성 있고 혁신적으로 개발되어 사회적 혜택을 증진하는데 있다. 인공지능의 사회적 활용을 위한 인공지능 윤리와 거버넌스, 데이터, 기술, 공공 및 민간부문 활용에 관하여 정책을 수립한다.(https://www.gov.uk/government/organisations/office-for-artificial-intelligence/about 2022년 3월 1일 최종접속).

57) 2015년 설립된 데이터과학 분야 정부와 13개 대학 협력 국립연구기관이다. 2017년 인공지능 분야 연구로 확장되었다.(https://www.turing.ac.uk/ 2022년 3월 1일 최종검색)

58) Office for Artificial Intelligence, *A guide to using artificial intelligence in the public sector,* 2020, 39면.

집단에 차별적 영향을 미칠 가능성을 고려하고, 설계와 실행의 단계마다 공정성 문제를 유의하며, 편향적 요소를 줄여야 한다. 셋째, 시민의 신뢰에 부응할 수 있어야 한다. 인공지능 기술의 안전성, 정확성, 신뢰가능성, 안정성과 건전성(safety, accuracy, reliability, security, and robustness)을 보장해야 한다. 넷째, 정당화 가능해야 한다. 설계와 실행 단계에서 투명성, 그리고 판단과 행태의 정당화 및 해석가능성을 우선순위에 두어야 한다.[59]

(2) 형사사법에서 인공지능기술 활용의 윤리적 기본틀

알란튜링 연구소의 형사사법체계내 인공지능 및 데이터기술 관련 윤리적 기본틀 구축(Building an ethical framework for data science and AI in the criminal justice system) 기획은 2019년 인공지능 윤리 및 안전 지침의 후속작업으로서 법무부(Ministry of Justice) 등 유관정부기관과 협업하에 특히 형사사법 분야 인공지능 기술활용 특화된 지침을 과제로 삼고 있다는 점에서 주목할 만하다. 이는 형사정책에서도 감시카메라나 드론, 금융사기 탐지알고리듬, 예측선제적 치안(predictive policing) 등 인공지능기술 도입과 활용이 확대될수록 부당한 감시와 침해에 대한 우려도 높아진다는 문제 때문이다.[60] 또한 시민의 안전과 권리보호를 위해 도입된 최신 인공지능기반 범죄예방기법이나 도구가 '스마트 치안(smart policing)'의 외관에도 불구하고 오히려 신뢰하기 어렵고 책임소재를 따질 수 없다면 그 자체 고위험 인공지능기술의 활용문제가 될 것이기 때문이다.
동 기획은 형사정책적인 인공지능기술 활용과 관련하여 제기될 윤리적 현안문제에 대처하고, 포용적이고 개방적인 관점에서 공적 활용될 수 있도록 보장하는 방안을 모색하기 위함이다.[61] 형사사법체계내에서 인공지능기반 자동화된 결정내용을 어떻게 설명가능하게 할 것인가, 신뢰성, 공적 책임성, 공정성과 같은 윤리적 가치를 어떻게 설계하고 실현할 것인지는 민간부문보다 오히려 더욱 중요한 문제이기 때문이다.
형사사법체계상 데이터과학 및 인공지능의 윤리적 기본틀의 일차적 목표는 형사사법기관 정책결정 및 판단책임자가 인공지능기술과 데이터과학의 혁신적 내용과 위험을 이해하는데 도움이 되는 윤리적 기본틀과 판단도구를 제공하는데 있다. 이를 위해서는 인공지능기반 기술이 개발·적용될 형사사법에 특유한 조직구조와 절차에 대한 이해, 해당기술의 최종사용자가 누구이며, 이러한 기술활용이 형사사법 직무에 어떠한 영향을 미치게 될지에 대한 검토가 필수적이다. 궁극적으로 인공지능 윤리가 요청하는 공적책임성(accountability)과 설명가능성(explainability)의 원리는 윤리적 가치에 기반한 형사사법기관간 소통을 근거짓고 동기지워주는 지침으로서, 편견이나 차별없는 공정한 조치의 실현가능하고 작동가능한 기반이 되어야 한다. 예컨대 인공지능기반 의사결정시스템이 설명가능한 절차를 통해 조정가능한 결과를 산출해 낼 수 있으려면, 형사사법기관의 거버넌스와 조직역량 또한 윤리적 기준에 부합될 수 있어야 한다.[62]

III. 고위험 인공지능기술의 활용과 형법적 통제

59) Office for Artificial Intelligence, A guide to using artificial intelligence in the public sector, 42-43면.
60) Peter Stone et al, Artificial Intelliegence and Life in 2030, 8면.
61) https://www.turing.ac.uk/research/asg(2022년 3월 1일 최종검색)
62) https://www.turing.ac.uk/research/research-projects/building-ethical-framework-data-science-and-ai-criminal-justice-system (2022년 3월 1일 최종검색)

유럽연합 등 고위험 인공지능기술에 대한 통제 논의는 공통적으로 인공지능기술의 활용이 시민의 안전과 권리를 침해하는 문제를 상정한다. 현실적으로 인공지능기술의 고위험이 구현되는 경우로서 형법적 제재의 필요성과 타당성까지 검토될 실제 문제지점은 지능형 기계(인공지능기반 로봇), 그리고 인공지능기반 자동인식·판단결정시스템이다. 인공지능 로봇과 시스템이 윤리적 규율을 위반하여 설계되거나 작동됨에 따라 시민의 안전과 권리를 침해한 경우, 사람이 이를 범죄적 목적으로 오용 내지 악용하거나 그 결과 침해에 이른 경우가 형법적 통제가 고려될 문제사안이 된다.

1. 인공지능기반 로봇의 위험

인공지능기반 로봇의 현실적 위험 유형은 산업용 로봇, 의료용 로봇, 자율주행운송 로봇, 성기능 로봇, 그리고 군사용 로봇 다섯 가지로 살펴볼 수 있다.

(1) 산업용 로봇

산업용 로봇의 공장 도입은 1960년대부터 시작되었고, 로봇의 '살인' 역사는 1979년 미국 미시건 주 포드 자동차 공장에서 시작된다. 기계식 팔 형태의 로봇이 일시정지하자 직접 부품을 가져오려던 노동자가 갑자기 다시 작동한 로봇에 부딪혀 사망한 사건이다. 이를 계기로 산업용 로봇은 안전펜스 등으로 격리되고, 사람의 개입이 필요하면 완전정지 후에 로봇 동작 영역에 들어갈 수 있도록 규정되었다. 현재는 사람과 협동해 안전하게 일할 수 있는 기계를 지향하는 협업형 산업용 로봇으로 대체되어 간다. 즉 일본에서 개발된 CR-35iA의 경우 contact stop 센서와 dual channel safety 소프트웨어를 장착해 로봇의 모든 방향에서 일어나는 접촉을 감지하고 사람과의 충분한 간격을 자동 모니터링할 수 있다. 이는 사람에 대한 피해방지에 초점을 맞춰 설계된 로봇의 보편화를 전망케 해준다. 또한 산업용 로봇이 초래할 산업재해가 최소화되고 사람과 활동공간을 공유하는 로봇에 대한 가장 중요한 문제는 안전보다 사람들의 인식이 될 것이다.[63]

산업용 로봇에 의한 생산공정은 노동인력의 효율적 대체뿐만 아니라 위험작업을 수행하여 산업재해를 줄일 수 있다는 측면도 있다. 산업용 로봇과 노동자가 협업하는 공정에서 안전설계 (safety by design)의 법적 의무화는 로봇의 작동소프트웨어와 구성장치가 작동의 효율성보다 로봇 작업환경내 사람의 안전을 우선하도록 하도록 하는 내용이 된다. 따라서 인공지능 로봇에서 안전설계의 형법적 책임은 제조물책임법상 '합리적 대체설계'와 구별될 것이다. 제조물책입법상 결함이란 해당 제조물의 제조상 · 설계상 또는 표시상 결함이 있거나 그 밖에 통상적으로 기대할 수 있는 안전성이 결여되어 있는 것을 말하는데, '설계상의 결함'이란 제조업자가 합리적인 대체설계를 채용하였더라면 피해나 위험을 줄이거나 피할 수 있었음에도 대체설계를 채용하지 아니하여 해당 제조물이 안전하지 못하게 된 경우를 말한다. (제2조 제2호 나목) 이러한 제조물 결함으로 생명 · 신체 또는 재산에 손해에 대해서는 배상책임이 부과된다.(동법 제3조 제1항) 또한 피해자가 ①해당 제조물이 정상적으로 사용되는 상태에서 피해자의 손해가 발생하였다는 사실, ②당해 손해가 제조업자의 실질적인 지배영역에 속한 원인으로부터 초래되었다는 사실, ③당해 손해가 해당 제조물의 결함 없이는 통상적으로 발생하지 아니한다는 사실을 증명한 경우에는 제조물을 공

63) David Hambling·Liron Gilenberg, WE: Robot, 13-14면.

급할 당시 해당 제조물에 결함이 있었고 그 제조물의 결함으로 인하여 손해가 발생한 것으로 추정한다. 다만, 제조업자가 제조물의 결함이 아닌 다른 원인으로 인하여 그 손해가 발생한 사실을 증명한 경우에는 그러하지 아니하다.(동법 제3조의 2)
이와 대조적으로 산업용 로봇의 하드웨어 및 소프트웨어상 안전설계 부재 내지 결함으로 인한 피해에 대해서는 안전설계의무 위반의 형사책임이 부과될 것이다. 또한 제조물책임법에서의 제조물 결함으로 인한 손해추정 규정과 대조적으로 검사가 정상적 사용상태에서 피해가 발생하였다는 사실을 증명한 경우 안전설계 의무위반을 추정케 하되, 로봇 설계제조업자가 제조물 결함이 아닌 원인으로 피해가 발생하였음을 증명한 경우에는 산업용 로봇 사용관리자의 안전관리 의무위반을 추정케 하는 방식이 될 것이다.

(2) 의료용 로봇

복강경 수술로봇 (Da Vinci Surgical System)은 심장판막 수술처럼 복잡하고 정교한 수술에 투입되어 의료사고나 합병증 위험을 줄일 수 있다. 이러한 로봇 수술은 외과의사에 의한 수술보다 비용이 높아서, 인건비절감을 위해 로봇을 도입한다는 통념과 반대되는 경우다. 물론 외과의사가 다빈치 수술로봇을 100% 통제하도록 설계되어 있다. 내시경 수술로봇 (Flex Robotic System)도 외과의사에 의한 원격제어 시스템이며 이제까지 어려웠던 종양제거수술도 가능하게 된다. [64]
미국 식약청(FDA)은 2000년 외과의사에 의한 상시 통제를 조건으로 사용을 허가했으며, 우발적 움직임 제어장치까지 설치되어 있다. 그러나 의사에 의한 추가시술을 필요로 하는 중한 부작용이나 의료사고 사망 사례도 보고된다.[65] 장기적 부작용 효과와 비용문제도 지적된다.[66]

(3) 성기능 로봇

성기능로봇은 장애인, 노인 등을 위한 반려로봇(companion robot) 또는 치료목적의 기능관점에서는 긍정적 가능성도 인정되는 반면, 인간의 존엄에 반한다는 반론도 맞선다. 영국의 1998년 인권법(Human Rights Act 1998)과 2010년 평등법(Equality Act 2010)에 따르면 장애인이 자신의 사적 공간에서 보통사람과 같은 행복을 추구할 권리를 제한하는 조치는 불법이라는 점을 고려하면, 권리의 문제가 될 수도 있다.[67] 이는 성행위가 기본적 인권의 하나이며, 성기능 로봇은 성행위 상대방에 대한 접근이 제한된 사람에 대한 일종의 해결수단이 될 수 있다는 인식이다.
그러나 성기능로봇에 대해서는 성폭행과 여성혐오적 폭력 실행 도구가 되어 여성의 성적 대상화와 폭력을 악화시킬 수 있다는 문제제기가 당연하다. 기계와 소통에 빠져 인간다운 상호공감과 감정교류를 상실하게 되리라는 더 근본적인 우려도 나온다.[68] 다만 인셀(incel)[69]의 성폭력범죄에 대한 방지책으로서 성기능 로봇의 가능성에 대한 주장도 있다.[70]

---

64) David Hambling·Liron Gilenberg, *WE: Robot,* 69-70, 89-90면.
65) https://www.drugwatch.com/davinci-surgery/ (2022년 3월 1일 최종검색)
66) Bec Crew, *Worth the cost? A closer look at the da Vinci robot's impact on prostate cancer surgery*, Nature 580, S5-S7,2020
67) Noel Sharkey et al, *Our Sexual Future with Robots, Foundation for Responsible Robotics*, 2017, 22-24면.
68) Jenny Kleeman, *Sex Robots & Vegan Meat*, 고호관 역, AI 시대, 본능의 미래, 2020, 113면.
69) 이른바 비자발적 독신주의자 남성으로 성적 좌절감을 여성혐오와 폭력으로 분출하며, 온라인 활동을 넘어 무차별 살인범죄로 나아가기도 한다.(Jenny Kleeman, Sex Robots & Vegan Meat, 83면.)

그러나 욕망과 권리는 다르며, 그 보장이나 해결수단의 의미 또한 달라야 한다. 무엇보다도 성폭력범죄에 대한 해결책은 못된다. 오히려 성폭력범죄의 증상에 더 가깝다. 개인욕구와 자기결정 침해여지 없이 전적으로 통제하고 지배할 수 있는 성적 대상을 찾는 자에게 성기능 로봇은 욕망을 해소하기보다는 촉발한다.[71] 아동형 성기능 로봇의 형법적 금지도 마찬가지 이유다. 소아성애자의 범죄를 방지하기보다는 오히려 범죄유발적이기 때문이다.[72] 특히 인간신체와 매우 유사한 로봇을 성폭력 범죄욕구의 대체물로 활용한다는 구상은 윤리적으로 사회적으로 수용하기 어렵다는 점은 분명하며, 범죄감소의 효과는 미미하고 불확실한데다가, 오히려 성폭력을 유발하는 위험한 경로를 터주는 일이 될 수 있다.[73]

미국의 경우 2021년 1월 연방 하원에 상정된 아동성착취로봇금지법안(Curbing Realistic Exploitative Electronic Pedophilic Robots Act 2.0)에 따르면 금지위반에 대하여는 5년 이하의 구금형이 부과된다. 또한 2019년 호주 연방의회에 상정된 아동성착취근절 개정법안(Combatting Child Sexual Exploitation Legislation Amendment Bill 2019)에 따르면 아동형체 성기능 인형 내지 로봇 제작·소지는 15년 이하의 구금형에 처해지는 중범죄다.[74] 그러나 이러한 법안들이 실제 법제화 단계에 이르지는 못하고 있는데, 아동형체 인형 내지 로봇은 그 자체 위험성이 있는 것이 아니라 도구에 지나지 않는다는 점에서 아동성폭력 내지 아동성착취에 대한 과도한 우려를 빌미로 위헌적으로 표현의 자유를 제한한다는 비판[75]때문인 것으로 판단된다.

한국의 경우 현행 청소년성보호법상으로 아동형체 성기능로봇은 아동성착취물[76]에 해당되지 않지만, 법취지에 비추어 보면 형법적 금지대상에 포함시킬 가능성이 높다. 그러나 성인형체 성기능로봇이 우리 사회 건전한 성풍속을 침해하므로 형법적 금지의 대상이 될 것인지 사생활의 자유 영역에 둘 것인지, 그리고 형법적으로 금지된 성폭력인 성매매의 경우 로봇 성매매[77]도 포함될 것인지는 별개의 판단을 요하는 문제다. 2021년 법원은 성인 섹스돌에 대해서 풍속을 해치는 물품에 해당하지 않는다고 보고, 매우 사적인 공간의 활동이므로 국가가 개입하지 않음으로써 자유를 실현해야 한다고 판단한 바 있다.[78] 또한 현행 성매매처벌법상 성매매알선행위는 성매매를 알선, 권유, 유인 또는 강요하거나 성매매 장소를 제공하는 행위(동법 제2조 제1항 제2호)인데 성매매 개념[79]상 성기능로봇은 포함되지 아니하므로 성매매근절목적의 성매매금지의 대상이 될 법적 근거가 없다.

---

70) Ross Douthat, *The Redistribution of Sex*, NYT 2018년 5월 2일자; Toby Young, *Here's What Every Incel Needs : a Sex Robot*, Spetator, 2018년 5월 5일자.

71) Jenny Kleeman, Sex Robots & Vegan Meat, 84-85면.

72) 김한균, 아동형체 '리얼돌' 성범죄화에 관한 영미국가 입법동향 비교 분석, 비교형사법연구 21(4), 2020, 109면.

73) Noel Sharkey et al, *Our Sexual Future with Robots,* 31면.

74) 같은 논문,109, 119면.

75) 같은 논문,110면.

76) 청소년성보호법 제2조 제5호. 5. "아동·청소년성착취물"이란 아동·청소년 또는 아동·청소년으로 명백하게 인식될 수 있는 사람이나 표현물이 등장하여 성교행위를 하거나 그 밖의 성적 행위를 하는 내용을 표현하는 것으로서 필름·비디오물·게임물 또는 컴퓨터나 그 밖의 통신매체를 통한 화상·영상 등의 형태로 된 것을 말한다.

77) 2018년 캐나다 성기능로봇 제작업체가 캐나다 토론토에 이어서 미국내 로봇 성매매업소를 시도했으나 불허되었다. ("Toronto's first sex doll brothel has been open for a year" https://www.blogto.com/city/2018/08/toronto-sex-doll-brothel/ ; "Sex robot brothel planned for Houston comes to a halt amid allegations owner misrepresented business" (https://abc13.com/sex-robot-brothel-sexbots-kinkys-dolls-galleria/4362686/ 2022년 3월 1일 최종검색)

78) "리얼돌, 음란 물품 아닌 성기구...법원, 수입 허용 판결" (조선일보 2021년 1월 25일자)

79) 성매매처벌법 제2조 제1항 제1호. "성매매"란 불특정인을 상대로 금품이나 그 밖의 재산상의 이익을 수수하거나 수수하기로 약속하고 성교행위나 유사성교행위를 하거나 그 상대방이 되는 것을 말한다.

(4) 자율주행운송 로봇

2016년 아마존이 시범도입한 배송용 드론(Amazon Prime Air)의 경우 상품배송의 자동화설정이 필요한데, 가장 현실적 문제는 인구가 밀집한 도시환경에서 안전하게 자율비행할 수 있는지 안전성 입증이다. 현재 인간 통제 시야를 벗어난 비행은 금지되고, 유인항공기 항로를 방해하지 않아야 한다.[80] 드론에 탑재된 센서로 장애물을 탐지하는 감지회피 시스템(sense-and-avoid system), 드론 기지국 역할의 비행선이나 배달 트럭, 고층건물 충전소 등 서비스 시스템 구축 등 관련된 문제[81]는 다양한 규제요인이 된다.

2017년 두바이 정부가 경찰로봇(Reem Dubai Police)과 함께 도입한 자율주행 경찰차(O-R3)는 비정상탐지 소프트웨어를 설치하여 범죄를 자동으로 감시하고, 용의자 추적을 위해 쿼드콥터(Quadcopter)를 발사할 수 있다. 쿼드콥터에 테이저건, 후추 스프레이를 장착하여 진압까지 가능한 기능도 시험단계다.[82]

이처럼 자율주행운송로봇은 자율주행차량부터 드론에 이르기까지 사람과 화물 운송용도로 일차 개발되지만 병기와 경찰로봇으로 전화할 가능성도 높고, 그 가능성은 집회시위진압이나 전투상황에서 인간 경찰과 병사와의 상호작용이 아닌 로봇과 대면하는 상황에서의 위험으로 나타날 수 있다.

(5) 군사용 로봇

드론이나 순항미사일과 달리 자동화 무기(Autonomous weapons)는 자율적으로, 즉 인간의 최종개입 없이 표적을 택하고 공격을 결정한다는 점에서 살상로봇(killer robot)이라 불리며 인공지능기술이 인류의 종말을 초래하리라는 경고가 가장 먼저 나오게 한 문제다. 2015년 전세계 인공지능 로봇과학자들이 서명한 공개서한은 가까운 장래 인공기능기반 무기가 군비경쟁의 수준으로 개발될 것을 예측하면서, 핵무기와 생화학무기와 마찬가지로 국제적 금지조약 필요성을 제기했다.[83] 2106년 유엔 전문가그룹은 자동화살상무기시스템 (lethal autonomous weapons systems)에 대한 국제적 논의를 시작했으며, 현행 특정재래식무기협약 (Convention on Certain Conventional Weapons) 체계하에서 특정 유형 무기 규제와 금지차원에서 논의된다. 그러나 자동화무기의 자동성(automated) 또는 독자성(autonomous) 개념규정에 있어서 표적설정과 감독에 대한 인간 개입정도에 대해서는 국가마다 입장이 갈린다. [84]

현실적으로 인공지능기술적 요소를 상당부분 구현하는 드론형태는 1995년 유고슬라비아 내전부터 투입되었고, 아프가티스탄 전쟁 당시 소위 테러리스트 표적을 탐지하게 되면서 자체미사일을 탑재하여 공격용으로도 사용되기 시작했다.[85] 원격조정된다는 점에서 자동화무기는 아니지만, 목표조준은 통신위치 등 정보인덱스에 근거하는 부분 자동화가 실현되어 있다.[86] 세계에서 가장 처음으로 그리고 유일하게 실전운용되는 군사용 로봇은 한국 DMZ에 배치되어 있는 SRG-A1 Sentry다. 표적을 감시하고 교전도 가능하다.

---

80) Federal Aviation Administration, *Package Delivery by Drone (Part 135)* (https://www.faa.gov/uas/advanced_operations/package_delivery_drone/ 2022년 3월 1일 최종검색)

81) David Hambling·Liron Gilenberg, *WE: Robot,* 93-94면.

82) 같은 책, 222면.

83) Future of Life Institute, Autonomous Weapons: An open letter from AI & Robotics Researchers, 2015 (https://futureoflife.org/2016/02/09/open-letter-autonomous-weapons-ai-robotics/ 2022년 3월 1일 최종검색)

84) House of Lords, Report of Session 2017-19, paras.342-343

85) The Intercept, *The Drone Papers* (2015.10.15) (https://theintercept.com/drone-papers/ 2022년 3월 1일 최종검색)

86) Thomas Powers & Jean-Gabriel Ganascia, The Ethics of the Ethics of AI, 38면.

카메라와 인식소프트웨어에 기반하여 감시영역에 들어오는 사람에게 암호를 묻고 올바른 암호를 감지할 수 있으며, 제스처 인식기능을 통해 항복의사도 확인할 수 있다. 기관총과 수류탄 발사기도 갖추고 있다. 현재는 고정모드에서 최종발사를 사람이 통제하지만, 이동모드를 선택하면 사람은 작동을 시작케 할 뿐 이동상태에서는 사실상 킬러로봇이 된다.[87] 인공지능기반 군사용 로봇의 투입은 상당한 부수적 피해도 이미 현실이지만, 대테러 수단이나 경찰용으로 전화될 수 있다.

2. 인공지능기반 자동인식 및 의사결정시스템의 위험

(1)생체정보인식시스템

공격용 드론이 출몰하는 분쟁지역이 아닌 일상공간에서 현재 가장 경계와 두려움의 대상인 인공지능기술이라면 생체정보인식 시스템(biometric identification systems)이다. 중국에서는 대도시 인구밀집환경에서 개인 동선을 추적 기록하는 기법이 소수민족 추적감시에 사용하고 있다는 것이다.[88] 보행신호위반 정도의 경범죄도 자동탐지하여 생체정보를 수집해 시민에 대한 "사회평판점수시스템 social credit system"이 구축될 가능성도 있다.[89] 행동예측이나 분류를 위한 안면인식(facial recognition)이나 감정분석 알고리듬에 대해서는 이를테면 신종 골상학(new phrenology)[90]이라는 비난마저 나오고 있다. 2020년 미국의 경찰폭력과 인종차별 사태와 맞물리게 되면 더욱 심각한 사회문제가 될 수 있다. 이에 따라 유럽과 미국, 중국 등에서 무차별적 안면인식 시스템의 감시에 대한 비판이 높아지고 있다.[91] 하지만 사회안전이나 테러방지를 명분으로 도입될 가능성이 높고, 일단 도입되면 활용 범위는 시민일반에로 확대될 위험성 또한 높다.[92]
이처럼 정부의 생체정보인식시스템을 통한 상시적 시민 감시와 일상 정보수집·축적은 위험으로부터 시민의 안전확보를 내세우지만, 권리침해라는 측면에서는 고위험 인공지능기술로서 통제가 필요하다. 고위험 인공지능의 위험에 대한 평가는 이처럼 기술적인 문제일 뿐만 아니라 사회적 문제이기도 하다.
유럽연합의 경우 생체정보인식시스템을 고위험 인공지능기술로서 엄격한 규제대상으로 규정한다. 공공장소에서 법집행기관의 사용은 금지대상이다. 극히 예외적으로 실종아동수색, 임박한 테러위협 탐지, 중대범죄 용의자나 범인 추적에 한하여 엄격한 요건하에 허용될 뿐이다. 즉 법원이나 독립된 판정기구에 의한 허용여부, 시간적 장소적 범위와 수집정보 범위 판단에 구속된다.[93]

(2) 자동화의사결정시스템

87) David Hambling·Liron Gilenberg, *WE: Robot,* 133-134면.
88) Paul Mozur, *One Month, 500,000 Face Scans: How China is Using AI to Profile a Minority*, NYT (2019.4.14.)
89) Rachel Botsman, *Big Data Meets Big Brother as China Moves to Rate Its Citizens,* WIRED (2017.10.)
90) Catherine Stinson, *Algorithms That Associate Appearance and Criminality,* American Scientist 109(1), 2021, 26면.
91) Thomas Powers & Jean-Gabriel Ganascia, The Ethics of the Ethics of AI, 34면.; Michael L. Littman et al, Gathering Strength, Gathering Storms, 36면
92) Thomas Powers & Jean-Gabriel Ganascia, The Ethics of the Ethics of AI, 38면.
93) European parliament, *Regulatory framework proposal on artificial intelligence* (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai 2022년 3월1일 최종검색)

자동화의사결정시스템 (automated decision-making system)은 심층신경망을 통해서 확보되는 고품질 데이터를 활용해 의사결정 기능을 수행하는 인공지능기술이다. 의료진단이나 대출심사, 신용카드 발급처럼 민감한 분야까지 포함해 많은 사회 영역으로 확대되고 있다. 일관성을 높이면서 비용은 낮추고 새로운 혁신적 해법을 통해 사회경제적 혜택을 가져올 수 있기 때문이다. 반면 오류 또는 차별적인 결과를 산출할 수 있고,[94] 나아가 인간의 오류를 강화하거나, 인간의 책임문제를 소거해 버린다[95]는 점에서 윤리적·법적 문제도 초래할 수 있다. 이러한 시스템 활용에 대한 시민의 신뢰상실로 이어지게 된다면, 오히려 사회적 기회비용 문제를 낳을 수도 있다. 자동화의사결정시스템 활용의 사회적 유용성과 윤리적 평가는 독자성(autonomy), 응용가능성(adaptability)과 확장성(scalability)의 문제가 된다.[96]

이는 자동화의사결정시스템 활용의 사회적 맥락을 통해 설명된다. 21세기 초기 경제침체기동안 공공부조, 저임금 일자리와 형사사법제도에서 정교한 데이터기발 자동화의사결정시스템이 확대되기 시작했다. 실제 도움을 필요로 하는 사람에게 서비스가 제공되어야 한다는 효율성 요구에 기반하여 예측 알고리듬, 위험평가모델, 자격판정 시스템은 신용평가, 마케팅, 양형, 재정, 공공프로그램 관리에서 일상화되었다. 하지만 특히 사회적 빈곤계층에게는 효율적 보호시스템이 아니라 세밀하고도 강화된 감시와 차별적 관리시스템으로 경험된다.[97] 자동화의사결정시스템은 정치적 책임이나 투명성을 기대하기 어려운 인권취약환경에서 우선적으로 도입되는데, 프라이버시보호나 보안조치 확보없이 개인정보를 수집하고 예측모형과 알고리듬을 통해 저신용자, 사회적 문제인물이라는 낙인을 찍는 결과에 이르기 때문이다. 이런 방식의 자동화 의사결정의 확산은 빈곤계층을 범죄자화하고, 차별을 심화시켜 사회안전망과 민주주의 제도역량을 오히려 약화시킨다.[98] 문제의 근본은 바로 알고리듬의 편향가능성에 있다. 인공지능 기계학습 데이터가 편향되면 부당하게 편향된 결정을 막기 어렵고, 데이터의 편향 방지를 법적 조치 역시 데이터 생성의 배경인 사회문화의 기존 편향 반영까지 막기는 더욱 어렵다. [99]

따라서 2018년 유럽연합 개인정보규정(General Data Protection Regulation)은 인공지능 기술에게 프로파일링을 포함한 개인에게 중요한 영향을 미치는 법적 효과를 낳는 자동화 과정을 기반으로 결정을 내릴 권한부여를 명시적으로 금지하며,[100] 인종, 정치적 견해, 종교적 사상적 신념, 노조가입과 같은 개인정보의 공개, 신원확인을 위한 유전자정보, 생체정보 처리, 보건데이터와 성적 데이터의 처리를 금지한다.[101] 또한 정보주체에게 자동화 결정에 대한 설명을 요구할 권리를 보장한다.[102]

인공지능기술의 개발과 활용은 독자적 의식과 판단결정이 가능한 시스템이 아니라, 인간에게 설명될 수 있고 책임을 물을 권한과 인간의 자율성이 보장되도록 설계함으로써 인간과 기계의 협업을 통해 인간역

---

94) D.Leslie, Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, Alan Turing Institute, 2019

95) G-Z Yang et al, The grand challenges of Science Robotics. Sci Robots 3(14), 2018

96) Jakob Mökaner · Maria Axente, *Ethics-based auditing of automated decision-making systems: intervention points and policy implications*, AI & SOCIETY, 2021 (https://doi.org/10.1007/s00146-021-01286-x)

97) Virginia Eubanks, *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor* 2018, 김영선 역, 자동화된 불평등, 2018, 26-28면.

98) 같은 책, 30-32면.

99) Stuart Russell, *Human Compatible: AI and the Problem of Control,* 2019, 이한음 역, 어떻게 인간과 공존하는 인공지능을 만들 것인가, 2021, 189-190면.

100) Art. 22. Automated individual decision-making, including profiling

101) Art. 9 Processing of special categories of personal data

102) Art. 14 Information to be provided where personal data have not been obtained from the data subject

량을 향상하는데 목표가 있다. 인공지능기술도 인류의 강점인 협업시스템의 일부로서 인간과 자동화의사결정시스템 사이의 명확한 소통 연결이 존재해야 한다.[103]

바로 이러한 요청이 공적 책임성, 신뢰가능성, 설명가능성, 투명성, 공정성과 같은 윤리적 가치에 정향된 인공지능기술이 논의되는 이유이며, 특히 고위험 인공지능기술의 윤리적 규율과 형사법적 통제는 인간중심의 인공지능기술 발전과 향유의 핵심내용이 되는 것이다.

103) Stuart Russell, Human Compatible,193면.; Michael L. Littman et al, Gathering Strength, Gathering Storms, 71면

< 기조발제 - 토 론 문 >

고위험 인공지능기술의 윤리적 규율과 형사법적 통제

조선대 법학과 이원상

I. 서론

인공지능 기술이 점차 보편화 되면서 사회의 다양한 분야에서 폭넓게 적용되고 있습니다. 문제는 인공지능 기술의 확산에 대응하는 입법정책이 충분히 펼쳐지지 못하고 있으며, 그로인해 법률의 제정도 더딘 상황입니다. 그로 인해 인공지능과 관련된 논의가 구체화되어 진행될 수 있는 터전이 마련되고 있지 않기에 여전히 일반적이고 추상적인 논의에 머무르게 됩니다. 더욱이 한국인의 성향에 대해 가장 잘 알고 있다고 할 수 있는 최근 소천하신 이어령 교수님께서는 한국인이 개별화에 집중하기 보다는 일반화에 집중하는 경향을 보인다고 하였습니다. 즉, 인공지능에 대한 법적논의와 법제의 더딘 걸음에 한국인의 일반화 성향으로 인해 한국의 인공지능에 대한 법적논의는 좀처럼 박스권을 탈출하지 못하고 있다. 그런 가운데 오늘 학회에서의 논의에서는 박스권을 탈출하고자 하는 열띤 노력들이 기울여진 것으로 보인다. 발제자께서는 인공지능 기술의 위험성을 제대로 이해하고 그에 대응하기 위한 윤리적 규율과 형사법적 통제의 필요성을 구체적이고 면밀하게 검토해 주셨습니다. 어려운 주제에 대해 풍성한 연구를 수행하신 발제자의 견해에는 전적으로 동의합니다. 다만, 논의를 조금 더 풍성하게 하기 위해서 위험과 윤리적 규율, 형사법적 통제로 나누어 토론을 진행하고자 합니다.

II. 인공지능의 위험

발제자께서는 먼저 인공지능으로 인한 위험에 대해 서술하고 계십니다. 인공지능으로 인해 발생하는 위험이 막연한 위험감정인지, 추상적인 위험인지, 구체화된 위험인지 아니면 이미 위험이 실현되어 침해로 이루어지고 있는지를 구분해 보는 것은 법적인 분석을 위해 상당한 의미가 있다고 판단됩니다. 그리고 해당 위험을 어떻게 통제할 수 있는지를 고민해 봐야 할 것입니다. 그러기 위해서는 인공지능을 특정 임무를 수행하는 인공지능, 범용 인공지능, 특이점을 넘어선 인공지능으로 나누어 볼 수 있을 것입니다. 특정 임무를 수행하는 인공지능으로는 인공지능 청소기, 인공지능 배달로봇, 인공지능 고객센터, 전쟁용 드론, 자율주행자동차 등을 예를 들 수 있을 것입니다. 해당 인공지능은 인간을 대신해서 특정 기능을 수행합니다. 범용인공지능은 SF 영화나 드라마에서 나오는 것과 같이 인공지능이 인간을 대신하여 인간이 할 수 있는 모든일을 할 수 있습니다. 특이점을 넘어선 인공지능은 인간의 능력을 넘어서 마치 신과 같은 능력을 갖게 된 인공지능을 말합니다. 특이점을 넘어선 인공지능에 대한 위험은 위험감정일 수 있습니다. 그에 반해 완전한 범용인공지능이 아직은 완벽히 개발되지 않았지만, 영화 ‘엑스마키나’나 ‘아이로봇’에

서 나오는 범용 인공지능의 위험은 추상적 위험으로 볼 수 있을 것입니다(물론 영화에서는 구체적 위험 및 침해를 하였습니다). 그에 반해 현재 현역으로 특정한 업무를 수행하고 있는 인공지능의 위험은 구체적 위험 또는 일부 위험이 실현되고 있다고 할 수 있을 것입니다. 결국 인간이 인공지능을 통제할 수 있는 영역은 특정 임무를 수행하는 인공지능에 한정되며, 범용 인공지능에 대한 통제는 인공지능의 특성상 통제가 거의 어렵고, 특이점을 넘어선 인공지능에 대한 인공지능에 대해서는 오히려 인간이 통제를 받게 될 수 있다고 생각합니다. 따라서 발제자께서는 개인적으로 인공지능에 대한 통제가 어느 정도까지 가능하다고 생각하시는 여쭙고자 합니다.

II. 윤리적 규율

인공지능 기술이 가지고 있는 위험에 대한 반응으로 인간들은 윤리적 인공지능을 개발하려고 노력하고 있습니다. 따라서 발제문에서도 밝히고 있는 것과 같이 미국과 유럽을 비롯한 많은 국가들이 윤리적인 인공지능을 개발하도록 윤리적 기본틀을 설정하고 있습니다. 그런데 윤리적 지침의 적용 대상은 인공지능을 개발하는 국가나 기업, 개발자입니다. 인공지능 개발에 윤리적 지침을 결합하여 윤리적인 인공지능의 결과가 나오기를 기대하고 있는 것입니다. 그런데 문제는 과연 윤리를 프로그래밍화 하여 인공지능에 반영할 수 있는지 여부입니다. 프로그램은 인간이 이해할 수 있는 이상을 할 수 없습니다. 인공지능이 윤리적인 인공지능이 되기 위해서는 윤리가 프로그래밍 되어야 되는데 인간도 윤리에 대해 제대로 이해하지 못하는 가운데 윤리적 규율이 인공지능에 어떻게 심어질 수 있을지가 문제입니다. 그에 따라 우리가 결정할 수 있는 방법은 결국 인공지능을 만드는 인간이 윤리적 가치를 가지고 만들도록 해야 한다는 것입니다. 즉, 문제는 인공지능을 창조하는 인간도 윤리적이지 않고, 더욱이 윤리를 프로그래밍화 할 수도 없으며, 설사 프로그래밍화 되었다고 하더라도 학습을 통해 발전하는 인공지능이 윤리적으로 작동할 수 있는지에 대해 여전히 의구심이 생기게 됩니다. 일부 챗본 등과 같은 사건에서도 그와 같은 문제점들이 나타나고 있습니다. 이에 발제자께 윤리적 규율이 과연 인공지능에 제대로 이식될 수 있다고 생각하시는지 여부를 여쭙고자 합니다.

III. 형사법적 통제

발제자께서는 형사법적 통제의 예로 산업용 로봇, 의료용 로봇, 자율주행운송 로봇, 성기능 로봇, 군사용 로봇 등으로 구분해서 서술하고 계십니다. 이처럼 로봇이라는 공통점을 가지고 있는 이유는 소프트웨어적인 인공지능이 침해할 수 있는 법익은 개인정보나 재산상 이익 등으로 한정될 수 있지만 하드웨어를 갖고 있는 인공지능은 인간의 모든 법익을 침해할 수 있기 때문일 것입니다. 그런데 그런 인공지능에 대한 형사법적 통제는 형사법 영역에서 전통적으로 논란이 되고 있는 위험형법에 대한 논의, 윤리의 범죄화에 대한 논의, 형법의 최후수단성, 단편성, 보충성에 대한 논의 등 여전히 치열한 결론지어지지 않고 있는 주제들이 연관될 수밖에 없을 것입니다. 발제자께서는 이미 이 분야에 대해 수많은 연구를 수행하셨고, 이 분야의 최고 전문가라고 할 수 있습니다. 그래서 발제자께서는 현재의 형사법체계와 이론으로 인공지능과 관련된 형사법적 통제

목적을 달성할 수 있다고 생각하시는지, 아니면 지금의 체계와 이론으로는 한계가 있기 때문에 패러다임을 전환하여 새로운 체계와 이론으로 나아가야 할지에 대해 고견을 여쭙고자 합니다.

IV. 결론

그동안 미국이나 유럽, 한국 등 일부 선진국가들에서 인공지능과 관련된 총론적 논의는 어느 정도 진전이 있었다고 생각됩니다. 이제는 실현된 인공지능 기술 및 현상에 대해 구체적이고 세밀한 논의가 진행되어야 할 것입니다. 오늘 학회에서 그와 같은 논의가 이루어질 수 있어서 매우 의미가 있었다고 생각합니다. 다만, 논의가 좀 더 진전되기 위해서는 입법적 노력이 함께 있어야 할 것입니다. 무엇보다 인공지능에 대한 학회의 논의가 입법으로 이어지고, 법률이 집행되고, 다시 피드백 되는 선순환 구조가 이루어져야 할 것입니다. 그를 위해 우리 학회에서도 지속적으로 관심을 갖고 논의의 장을 마련해야 할 것입니다. 이렇게 귀한 자리에서 토론을 할 수 있도록 배려해 주신 학회회원님들과 학회장님께 깊은 감사를 드립니다.

세션 I

< 발 표 문 >

인공지능재범위험성평가도구의 도입필요성과 규범적 문제점
-보안처분의 판단과정을 중심으로-

장진환

I. 서론

1. 인공지능의 확장과 형사사법

인공지능에 대한 관심과 열기가 뜨겁다. 인공지능이 4차산업혁명의 핵심으로 부각되면서 전 세계적으로 인공지능에 대한 경쟁력을 선점하기 위한 전쟁이 이미 곳곳에서 진행 중이다. 우리나라도 이에 질세라 첨단IT기술을 통해 쌓아 올린 국가경쟁력을 이어가기 위해 인공지능과 관련해서라면 국가차원의 전방위적인 지원을 하고 있는 실정이다. 이러한 흐름만 놓고 보면, 나라마다 속도의 차이가 있을 뿐, 인공지능이 우리 삶의 전반에 걸쳐 광범위하게 적용되는 것은 결국 시간문제로 보인다. 이는 형사법 영역에서도 예외가 아니다. 범죄의 예방과 수사, 공판 과정에 이르기까지 여러 영역에서 미국을 중심으로 이루어지고 있는 제도들이 이미 우리나라에서 수많은 문헌을 통해 소개되고 있고, 심지어 몇몇 프로그램은 이미 시범운영 중에 있다.[1] 비록 지금은 그 활용도가 크지 않지만, 경쟁력제고를 위해 효율성과 비용절감을 무엇보다 중요시하는 사회전반적인 흐름과 인공지능을 비롯한 첨단 과학기술에 대한 투자비중과 국민적 호응도를 보면 앞으로 인공지능이 형사정책에서 차지하는 비중은 다른 영역에서처럼 점점 높아져 갈 것으로 예측된다. 이렇게 인공지능이 형사사법영역에서 확대 적용되는 것이 장려되어야 하는지 아니면 지금이라도 이러한 흐름을 단호하게 막아서는 결기가 필요한지에 대해서는 일괄적으로 평가할 수 없다. 이는 추상적인 차원에서 논의될 성격이 아니라, 개별 영역에서 논의의 대상이 되는 법제도 및 법 규정의 특수성을 기반으로 인공지능의 결합으로 인한 기대되는 효과, 문제점, 그리고 이러한 문제점의 보완 가능성 등을 종합적으로 고려하여 판단할 문제이기 때문이다. 따라서 이와 관련된 개별적인 영역에서의 구체적 검토가 요구된다.

2. 보안처분영역에서의 인공지능을 이용한 재범위험성예측도구의 도입가능성

본 글에서 관심을 가지는 영역은 인공지능을 이용한 재범예측 분야이다. 이 영역은 기본적으로 통계를 바탕으로 한 예측프로그램이 중심이 되고 있는 만큼, 빅데이터를 기반으로 하는 인공지능이 결합되기 용이하다. 따라서 그 어떤 다른 영역보다 인공지능의 적용을 둘러싼 논의의 의미가 크고, 이미 관련연구가 활발히 이루어지고 있는 분야이다.[2] 인공지능을 이용한 이러한 재범예측영역은 세부적으로 보면, 그 분야가 매

1) 현재 인공지능이 우리나라 형사사법에서 어떻게 적용되고 있는지와 앞으로의 발전방향에 대한 소개 문헌으로는 김웅재, 형사절차에서 인공지능 알고리즘의 활용가능성과 그 한계, 서울대학교 기술과 법센터, LAW & TECHNOLOGY, 제16권 제4호, 2020년 7월, 3-23면; 김기범·이관희·김일권·주성빈, 빅 데이터 환경에서의 범죄예측, 범죄 빅데이터를 활용한 범죄예방시스템 구축을 위한 예비연구(I), 한국형사정책연구원, 2014, 189-221면; 양주아·양문승, 경찰의 과학치안 활동 및 전망, 한국경찰연구 제16권 제4호, 2017, 147-168면; 윤상연·신상화 범죄 예측의 현재와 미래—한국형 범죄예측 시스템 개선 도입에 대한 제언, 한국공안행정학회보 2018, 제27권 3호, 248-272면.

2) 미국에서는 인공지능 재범예측도구인 컴퍼스(COMPAS)를 중심으로 많은 논의들이 이루어지고 있고, 국내에서도 많은 선행 연구들이 이러한 미국의 논의를 소개하면서 인공지능 재범예측도구가 가지는 편향성 문제, 투명성 문제 등의 문제점들을 법적인 쟁점과 연결시켜 다루고 있다. 이에 대한 대표적 문헌으로는 이병규, AI의 예측능력과 재범예측알고리즘의 헌법 문제- State v. Loomis 판결을 중심으로, 공법학 연구, 한국비교공법학회, 2020, 제21권 제2호, 169-191면; 김웅재, 형사절차에서 인공지능 알고리즘의 활용가능성과 그 한계, 서울대학교 기술과 법센터,

우 광범위하고 다양하다.[3] 예를 들어 과거범죄를 분석한 데이터를 기반으로 재범의 장소와 시간을 예측하는 프로그램도 있고, 특정 환경적 요인들 및 사회관계적 요인들을 바탕으로 장래 범죄를 저지르거나 범죄의 피해자가 될 위험이 높은 사람들을 특정하여 이들에 대한 주의 및 예방교육을 시도하고자 하는 프로그램도 존재한다.[4] 이하에서는 인공지능의 도입 및 적용과 관련하여 다양한 재범예측 분야 가운데서도 보안처분의 선고 및 해제와 관련된 영역에서 이루어지는 재범위험성 평가에 대해 다루어 보고자 한다. 보안처분은 기본적으로 형벌의 집행을 종료한 시민을 대상으로, 미래에 특정범죄를 다시 범할 위험성이 높다는 이유만으로 부과되는 형사제재이다. 따라서 사람에 대한 재범위험성 평가를 전제로 하고, 형벌에 못지않은 심각한 기본권 침해를 수반하기에 매우 엄격한 법치국가적 정형성과 정확한 재범위험성 예측이 요구되는 영역이다. 이와 관련해 인공지능은 한편으로는 기계가 가지는 특성으로 인해 재범예측의 객관성 및 정확성을 높일 수 있다는 기대를 받고 있지만, 또 다른 한편으로는 판단과정의 투명성과 설명가능성 등의 새로운 규범적 문제도 야기한다는 비판을 받고 있는 만큼, 보안처분 영역에서 인공지능의 활용가능성과 한계에 대해 검토해보는 것은 의미가 있다고 판단된다.

3. 논의의 지점

이를 위해 구체적으로 검토하고자 하는 지점은 크게 2가지이다. 첫 번째는 보안처분선고와 관련해 효율성과 정확성 측면에서 지금의 재범위험성평가 방식대신에 인공지능을 이용한 재범위험성도구를 도입할 필요성(실익)이 있는지의 문제이다. 재범예측의 정확성문제는 보안처분제도 정당성논쟁의 핵심쟁점 가운데 하나이므로 인공지능재범위험성예측도구 활용에 있어 항상 논의될 것으로 예상된다. 그러나 이에 대해서 아직 직접적인 판단을 내린 연구결과를 찾아보기 힘들어 현재의 재범위험성평가도구와 인공지능재범위험성예측도구의 비교분석을 통해 대략적이나마 이에 대한 마 답을 찾고자 한다. 두 번째는 효율성이나 정확성 측면에서 설사 인공지능을 이용한 재범위험성평가도구가 보안처분의 판단절차에 도입 될 실익이 있다고 하더라도, 규범적인 측면에서 문제가 제기될 수 있는 부분에 대한 검토이다. 아무리 필요성이 있다고 하더라도 규범적인 문제점이 있으면 도입 및 활용은 불가능하기 때문이다. 이에 대해서는 여러 가지 논점들이 존재할 수 있지만 본 글에서는 지면의 한계 및 기존연구와의 차별성을 위해 ①인공 지능에 의해 도출된 재범위험성 결과를 공판과정에서 사용할 수 있는지, ②인공지능에 의한 재범위험성평가과정을 의무규정으로 입법해야 하는지, ③인공지능에 의한 재범위험성 평가결과에 대한 법관이 구속되어야 하는지에 대해 검토하겠다.

II. 인공지능 재범위험성평가도구의 도입 필요성에 대한 검토.

1. 재범예측 방법의 종류 및 역사적 발전과정

인공지능 재범위험성평가도구의 도입 필요성을 검토하기 위해서는 비교 대상이 되는 인공지능 재범위험

LAW & TECHNOLOGY, 제16권 제4호, 2020년 7월, 3-23면; 한애라, 사법시스템과 사법환경에서의 인공지능 이용에 관한 유럽 윤리헌장의 검토 -민사사법절차에서의 인공지능 도입 논의와 관련하여 -, 2019, 38-79면; 주현경, 정채연, 범죄예측 및 형사사법절차에서 알고리즘 편향성 문제와 인공지능의 활용을 위한 규범 설계, 2020, 제27권 제1호, 115-162면.

3) 다양한 재범예측기술 활용현황에 대한 설명으로는 Angele Christin et al.,“Courts and Predictive Algorithms”, Data & Civil Rights: A New Era of Policing and Justice, Data & Society, 2015, 2-4면.

4) 이에 대한 자세한 소개는 Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, John S. Hollywood, Predictive Policing -The Role of Crime Forecasting in Law Enforcement Operations (Santa Monica, CA: RAND Corporation, 2013; 국내 문헌으로는 노성훈, 시공간 분석과 위험영역모델링을 활용한 범죄예측모형의 예측력 검증, 형사정책연구 제26권 제3호, 2015; 정진성 “주요 외국의 빅데이터 기반 범죄예방시스템 운영현황 및 성과 분석”, 범죄 빅데이터를 활용한 범죄예방시스템 구축을 위한 예비연구(II), 2015 등.

성평가도구와 현재의 평가도구의 특징파악이 필수적이다. 이와 관련해 우선 재범위험성 예측방법의 종류 및 전반적인 발전과정에 대해 살펴볼 필요가 있다. 현재의 재범예측방법은 과거의 재범예측 방법이 가지는 단점을 보완하기 위해 노력하는 과정에서 도출되어 왔기 때문에 각 방법들 사이에는 밀접한 관련성이 존재하고, 이는 인공지능을 이용한 예측방법이 기존의 예측방법과 어떤 점에서 유사하고, 또 어떤 점에서 차이가 있는지를 비교분석하는 데에 있어 중요한 단초를 제공할 수 있기 때문이다.

가. 직관적 방법

최초의 방식으로 인간행동에 대한 이른바 일상이론(Alltagstheorie)[5]을 바탕으로 재범위험성이 높거나 적다는 것을 예측하는 방법을 말한다. 말 그대로 직관적 방법이므로 평가자가 일상경험을 통해 습득된 사실이 평가기준이 된다.[6] 예를 들어 평상시 욕설을 자주 하는 사람이 범죄를 저지른 것을 많이 목격한 사람은, 평가 대상자가 욕설을 한다면 이는 재범가능성이 높다고 판단하는 형태이다. 이러한 직관은 사람마다 통일되어 있지 못할 뿐 아니라, 그 시대의 경험 상황을 전제로 한다는 한계를 가지고 있다. 특히 이 방법은 단지 판단자의 직관에 의존하기 때문에 과학적이지 못하고, 어떤 결과까지 도출되는 과정에 대한 설명가능성측면에서도 설득력이 부족하다는 평가를 받는다.[7]

나. 임상적 방법

임상적 방법은 직관적 방법에 의한 문제점을 극복하고자 심리학자나 정신과 의사 등의 전문가가 당사자를 광범위한 영역에 걸쳐 관찰하고 일반적인 범죄학적 지식을 바탕으로 예상되는 사안에 대한 위험성 요인 추출하여 이를 기반으로 재범가능성을 평가하는 형태를 말한다.[8] 예를 들어 당사자가 인격성 장애나 정서결핍 등의 정신적, 심리적 질병을 겪고 있다고 판단되면, 이러한 질병적 특징들이 범죄학적 지식에 비추어 볼 때, 얼마나 큰 위험 요소가 되고, 장래에 어떻게 발현될 것인지를 예측하여 최종적인 재범위험성을 판단하는 형태이다. 이때 중요한 것은 단지 특정 장애 혹은 질병을 겪고 있다는 사실만으로 재범위험성이 바로 추정되는 것이 아니라, 그러한 장애 혹은 질병 현상과 특정범죄의 발생요인이 얼마나 직접적으로 연관이 있는지 여부에 따라 재범위험성이 최종적으로 판단된다는 점이다.[9] 이러한 방법은 전문가가 대상자 개개인마다의 특징을 자세하게 살펴 판단하기 때문에, 위험성 평가과정에서 각 대상자가 가지는 개개인의 고유한 특징들이 잘 반영될 수 있고 판단의 근거를 설명할 수 있다는 장점이 있다. 그러나 모든 과정이 오로지 전문가의 주관적 판단에 맡겨져 있으므로, 전문가마다 대상자에 대한 평가가 다를 수 있다는 문제점도 동시에 존재한다.[10]

다. 통계적 방법

통계적 방법은 임상적 방법이 가지는 주관성의 문제점을 극복하고자 처음부터 재범위험요소를 분류해 놓

5) 일상이론이란 학문적으로 논증되거나 검증되지 않았지만, 개인적으로, 또는 사회적으로 통용되는 상식 수준의 주장을 가리키는 사회학 용어이다. 대부분의 사람들이 그렇다고 생각하지만, 이는 단지 그 당시 사람들의 생각일 뿐, 당연히 사실과 다를 수 있다.

6) Dessecker, Gefährlichkeit und Verhältnismäßigleit, S.196; Dahle, Handbuch der forens. Psychiatrie, 1-26면.

7) Schumann, Prognoseentscheidungen in der strafrechtlichen Praxis, 34면 이하; Meier, Kriminologie, § 7 Rn.39.

8) Eisenberg, Kriminologie, § 21 Rn.17.

9) Boetticher, Dittmann, Nedopil, Nowara, Wolf: Zum richtigen Umgang mit Prognoseinstrumenten durch psychiatrische und psychologische Sachverständige und Gerichte, NStZ 2009, 478면.

10) 같은 분야의 전문가더라도, 전문가의 지식, 참고하는 문헌 및 성향 등에 따라 판단결과가 다를 수 있다. 이에 대한 설명으로는 Meier, Kriminologie, § 7 Rn.31.

은 공통된 위험평가표를 만들고, 이에 따라 대상자의 특징을 평가하여 재범위험성을 판단하는 형태이다. 위험성 평가표는 범죄학적 지식 및 다수에 대한 과거의 범죄 관련 통계자료를 이용하여 만들어지고 전과사실, 가족관계, 직업유무 등 각 세부 항목별로 점수화가 되어 있다. 이러한 표에 대상자의 정보를 대입하면 최종점수가 나오고 그 점수가 특정 점수 이상이면 재범위험성이 있다고 판단된다. 이러한 통계적 방법은 평가자가 누구인지에 따라 달라지지 않고 평가과정이 투명하다는 장점을 가지고 있다.[11] 그러나 평가를 받는 대상자의 개별적 특징이 명확히 반영되지 못한다는 비판도 받는다.[12] 왜냐하면 위험점수표는 어디까지나 일반적인 평균의 값으로, 대상자에 따라서는 이러한 평균적인 특징과 차이가 나는 부분이 존재할 수 있기 때문이다.

라. 혼합적 방법

임상적 방법과 통계적 방법을 합친 것이다.[13] 이는 기존의 방법들이 가지는 특징과 문제점을 최대한 보완하고자 나온 방법으로, 현재 전 세계적으로 널리 사용되고 있는 형태이다. 앞서 살펴본 3가지 방법(직관적 방법, 임상적 방법, 통계적 방법)의 정확성에 대한 경험적 연구결과들에 따르면 어떤 하나의 방법이 다른 방법들보다 유의미할 정도로 높은 성과를 보인다고는 평가할 수는 없지만,[14] 그나마 직관적 방법보다는 임상적 방법과 통계적 방법이 아주 미세할 정도로 재범예측의 정확성면에서 우수하다고 보고되고 있다.[15] 이를 근거로 임상적 방법과 통계적 방법의 장점을 합쳐 정확성을 높이자는 움직임이 일어났고, 그 결과 나온 것이 혼합적 방법이다.[16] 이를 구체적으로 살펴보면, 우선 과거 통계적 방법에 사용되었던 각 항목 영역을 보다 세분화하고 항목별 판단요소들의 숫자를 훨씬 늘렸다. 예를 들어 판단 영역을 생애주기별 특징, 임상적 특징, 미래의 예상되는 생활형태의 특징 등으로 세분화하고, 각 영역에 따른 하위 항목들의 표를 만들어 과거보다 더 구체적으로 점수화하였다.[17] 이는 보다 다양한 범죄요인의 특성 및 대상자 개개인의 특별한 요소를 고려하겠다는 의미이다. 또한 이에 그치지 않고 이러한 위험판단 요소들을 일반인이 아닌 해당 전문가가 대상자 개개인의 특징에 맞도록 다시 한 번 종합평가하도록 하고 있다. 따라서 단지 통계적인 점수 결과가 일정 점수 이상을 획득하여 재범위험성이 높다고 추정되더라도, 전문가가 통계에서 미처 드러나지 않은 대상자의 특이점을 발견하였다거나, 개별 사건에서 특별한 이유로 인해 재범위험성 통계조사가 신빙성이 떨어진다고 판단하는 경우 이에 의존하지 않고 재범위험성이 높지 않다는 평가를 내릴 수도 있다.[18] 이는 재범위험성 평가가 자칫 지나치게 통계 중심의 일률적 평가로 흐르는 것을 막고, 대

---

11) BGH StV 2008, 301-302면.

12) Sigmund, Probleme und wissenschaftliche Methoden bei der Prognose von Gefährlichkeit, Freilaw, Strafrecht 2면.

13) 혼합적 방법 및 통계적 방법의 명칭과 관련하여 정리되지 않은 부분들이 존재한다. 우리나라 문헌들에서는 혼합적 방법을 통계적 방법 혹은 보험계리적 방법이라고 표현하는 경우가 많다. 이는 내용을 오해해서 잘못 표현한 것이 아니라, 서양을 중심으로 처음 학자들이 재범위험성판단의 발전과정을 분류하여 소개하는 과정에서 학자 각자의 기준과 강조점이 다른 점에서 기인한다. 즉, 어떤 문헌에서는 통계적 방법이라고 하면 위에서 소개한 것처럼, 처음 과거의 통계적 방법을 지칭하지만, 다른 일부 문헌에서는 이 글에서 말하는 혼합적 방법을 개량된 통계적 방법 혹은 보험계리적 방법이라고 지칭하고 있다. 우리나라의 문헌들에서도 이러한 다양한 서양의 문헌들 가운데 어떤 문헌을 참고하여 소개하는지에 따라 명칭이 달라지고 있다. 하지만 중요한 것은 명칭이 아니라 내용인데, 내용적인 측면의 발전과정에서는 특별히 유의미한 이견이 존재하지 않기 때문에 어떤 분류방식과 명칭을 택하더라도 큰 문제가 없다고 생각된다.

14) Eisenberg, Kriminologie § 21 Rn.39이하.

15) 이에 대한 자세한 소개는 Feltes, Rückfallprognose und Sicherungsverwahrung - die Rolle des Sachverständigen. In: Strafverteidiger 2000, 281-286면.

16)Boetticher, Dittmann, Nedopil, Nowara, Wolf: Zum richtigen Umgang mit Prognoseinstrumenten durch psychiatrische und psychologische Sachverständige und Gerichte, NStZ 2009, 478면.

17)Nedopil, Forensische Psychiatrie: Klinik, Begutachtung und Behandlung zwischen Psychiatrie und Recht, 2007, 293면 이하.

18)Boetticher, Dittmann, Nedopil, Nowara, Wolf: Zum richtigen Umgang mit Prognoseinstrumenten durch

상자 개개인의 가지는 개별적인 특징들을 전문가가 자세하게 관찰하여 최대한 반영하겠다는 의도이다. 따라서 혼합적 방법은 결국 발전된 통계적인 측면(보험계리적 평가)을 기반으로 보다 구조화된 임상전문가의 평가가 이루어지는 것을 의미한다.

2. 현재의 운영현황

현재 우리나라는 다른 나라들과 마찬가지로 가장 발전된 방법이라고 평가받고 있는 혼합적 방법을 사용하고 있다. 전 세계적으로는 미국과 캐나다를 중심으로 Psychopathy Checklist Revised (PCL-R)[19], Static-99[20], Historical Clinical Risk Management-20 (HCR-20)[21], 와 같은 범죄의 영역별로 다양한 재범위험성 평가도구가 개발되어 실무에서 사용되고 있고, 우리나라도 이를 기본 바탕으로 하여, 우리나라 현실에 맞는 요소들을 가미하거나 변경한 독자적인 재범위험성 평가도구를 만들어 사용하고 있다. 예를 들어 성폭력 범죄자에 대한 청구전조사에서는 국내에서 만들어진 재범위험성평가 도구인 K-SORAS가 실시되고 있고, 평가결과 위험성 수준이 '높음'에 해당할 경우나 피해자가 2명 이상, 또는 16세 미만 아동인 경우 등에는 PCL-R을 추가로 실시하도록 하고 있다. 이외에 살인이나 미성년자 유괴범죄자에게는 PCL-R을, 기타범죄자에게는 KORAS-G를 실시하고 중요 사건일 경우에는 PCL-R을 추가로 실시할 수 있도록 규정하고 있다. 그리고 이러한 재범위험평가도구 결과를 기반으로 임상전문가가 각 사례에 대한 재범위험성을 상, 중, 하로 평가한 종합평가보고서를 작성하여 검사나 법관에게 제출하는 형태를 취하고 있다.

3. 인공지능을 이용한 재범위험성 예측

가. 기본원리

형사법 영역에서의 인공지능의 활동은 크게보면 ①빅데이터 수집à②빅데이터 해석à③프로파일링(상관관계 패턴 등의 확정)à④구체적 사건 적용à⑤사후관리순으로 이루어진다고 볼 수 있다. 재범위험성평가 분야도 크게 다르지 않다. 우선 범죄자의 과거전과횟수, 범죄의 특징, 가족관계, 직업유무, 질병력 등등 매우

---

psychiatrische und psychologische Sachverständige und Gerichte, NStZ 2009, 478면.

19) PCL-R은 정신질병의 특성에 집중하여 재범위험성을 평가하는 도구로 사이코패시 체크리스트(Psychopathy Checklist-Revised)의 줄임말이다. 한국판 PCL-R은 미국의 정신의학자인 클레클리(Hervey M. Cleckley)가 정신병질의 특성을 정리하여 발표한 1941년의 연구를 바탕으로 캐나다 심리학자인 헤어(Robert D. Hare)가 개발한 평가도구를 한국어로 번안한 것이다. 헤어는 1980년에 사이코패시 평가도구를 개발하였고 1991년에 이를 수정한 버전을 출판하였으며, 한국에서 번역한 것은1991년 수정버전이다. 이러한 내용의 자세한 소개에 대해서는 유진, 위험한 범죄자와 위험관리체제의 성격, 176면.

20) Static-99는 성범죄자 재범위험성 평가도구로 캐나다의 핸슨(R. Karl Hanson)이 개발한 RRASOR(Rapid Risk Assessment for Sexual Offense Recidivism)와 영국의 쏘튼(David Thorton)이 개발한 SACJ-Min(Structured Anchored Clinical Judgement-Minimum)에서 사용하는 위험평가요소를 결합하여 만들어졌다. 우리나라에서는 이를 바탕으로 우리나라 실정에 맞게 변환된 K-SORAS(Korean Sex Offender Risk Assessment Scale)를 개발하여 사용하고 있다.

21) HCR-20은 폭력위험성을 판단하는 평가도구로, 1933년 Webster 등이 주도하여 캐나다 밴쿠버에서 개발된 4세대 HCR-20 위험성 평가가 주로 사용되고 있다. 국외에서는 가석방 심사, 교정 내 처우 분류, 치료감호소, 보호관찰, 우범자관리, 보호관찰, 사법기관의 폭력위험성 평가 등에서 폭넓게 사용되고 있으나, 우리나라에서는 아직 사용되고 있지 않다. 이를 국내에 자세한 소개한 문헌에 따르면 HCR-20는 소위 보험 계리적 도구(actuarial instrument)가 아닌 구조화된 전문가 판단 모델(structured professional judgement model)을 바탕으로 만들어져 근본 취지는 규범 정립적 접근에서 벗어나 임상가가 개별기술적인 방식에서 폭력 위험성을 평가 하고 이를 바탕으로 사례를 개념화하여 향후 펼쳐질 수 있는 위험성 시나리오를 발견하는데 도움을 주는데 있다고 한다. 따라서 PCL-R, Static-99 등의 다른 계리적 도구처럼 각 위험요인을 점수를 합산하여 등급을 분류하고자하는 것은 본 도구의 근본적인 목적을 훼손하는 것이라고 하고 있다. 이에 대해 자세한 설명은 서종한, 조병철, 김희송, 김경일, 전문가판단모델(SPJ) 기반 폭력위험성평가(HCR-20) 버전 3 타당화 연구: 재범예측에서 예방으로의 패러다임 전환, 형사정책연구 제29권 제1호, 2018, 309-336면.

광범위한 분야에 걸쳐 빅데이터들[22]이 수집된다. 이러한 빅데이터는 그 자체로는 특별한 의미를 지니기 힘들기 때문에 인공지능을 이용하여 이를 분석하는 과정이 필요하다. 이때 기반이 되는 분석기술의 기본은 숫자나 확률을 이용한 통계적 기법이다. 이러한 통계적 기법을 기반으로 재범위험성평가 프로그램은 알고리즘[23]에 따라 수집된 빅데이터들을 분석하여 범죄예측과 관련한 유의미한 상관관계를 획득해 내고, 어떤 조건들이 충족이 되면 위험성이 높다고 판단할지에 대한 판단요소들을 설정한다. 그리고 해당대상자의 자료들을 이에 대입하면 해당 대상자의 재범위험성점수(확률)가 얼마인지를 계산해낸다.[24] 또한 사후적으로 판단의 오류가 있었던 경우에는, 예를 들어 위험하지 않다고 판단하여 보안처분을 부과하지 않고 사회에 내보내었는데, 당사자가 다시 범죄를 저지른 경우 등은 새로운 데이터로 수집되어 기존의 알고리즘에 반영된다.

나. 인공지능의 핵심기술인 머신러닝과 딥러닝

일반적으로 인공지능에서 사용되는 알고리즘은 설계자의 의도가 반영된다. 왜냐하면 알고리즘과정에서 이루어지는 의사결정 단계, 즉 기준 설정, 훈련 데이터의 분류 및 선정, 의미론(semantics) 및 해석, 포함 및 배제되는 정보의 선택 등의 과정에서 수집된 빅데이터의 존재만으로는 특정기준을 설정할 수 없기 때문이다.[25] 그런데 이러한 알고리즘 과정에서는 시간이 지나고 판단데이터가 많이 쌓임에 따라 인공지능이 단순히 프로그램된 지시사항을 수동적으로 수행하는 것이 아니라 새로운 데이터를 통해 스스로 학습하고 결과물을 산출하는 과정도 포함되어 있는데, 이를 머신러닝(기계학습)이라고 한다. 머신러닝은 정형적이고 구조화된 데이터뿐만 아니라 대량의 비정형적·비구조적 데이터를 종합적으로 처리하면서 알고리즘을 발전시켜나가므로 빠르게 학습 가능한 규칙 및 새로운 지식을 자동적으로 추출해내어 기존에 설계된 프로그램의 결과값과 다른 판단이 가능할 여지를 만든다. 머신러닝은 지도학습, 비지도 학습, 강화학습 등의 방법을 사용하여 학습을 하는데,[26] 이와 관련해 딥러닝이란 개념을 살펴볼 필요가 있다. 딥러닝은 머신러닝의 하위개념으로 자가 학습 면에서 기본머신러닝모델보다 한 단계 더 진화된 개념이다.[27] 기본 머신러닝모델

---

22) 빅데이터란 단지 데이터가 많다는 것만을 의미하지 않는다. 인공지능에서 말하는 빅데이터는 크게 3가지 측면(규모, 다양성, 속도)에서 기존의 데이터들과는 다른 특징을 보인다. 규모의 측면에서 빅데이터는 모아지는 데이터의 크기가 테라바이트(TB) 또는 페타바이트(PB)에 이르고 있고, 다양성의 측면에서 정형적、구조적 데이터가 아니라 동영상, 음악, 사진, 위치정보, 각종 로그기록 등 그 내용이 달라 통일적인 구조로 처리하기 쉽지 않은 비정형 내지 비구조적 데이터이며, 속도의 측면에서 매우 빠른 속도로 실시간으로 계속 수집되는 특징을 가진다. 이에 대한 설명은 김혜정, 형사제재에 있어 빅데이터를 이용한 재범위험성예측의 활용 프레임 재정립을 위한 시론적 연구, 형사법의 신동향 통권 제66호, 2020 124-125면.

23) 알고리즘은 어떤 문제를 해결하거나 특정 결과를 얻어내기 위해 착수하는 일련의 단계들로 정의될 수 있다. 그 구체적인 단계들은 우선화(prioritization), 분류(classification), 연관(association), 그리고 필터링(filtering)으로 이루어지고, 이러한 연쇄적인 단계들을 거쳐 최종적인 의사결정이 도출된다. 주현경, 정채연, 범죄예측 및 형사사법절차에서 알고리즘 편향성 문제와 인공지능의 활용을 위한 규범 설계, 2020, 제27권 제1호, 120면 참조.

24) 인공지능 알고리즘에 의한 위험성 예측도구는 방대한 데이터를 바탕으로 수백 가지의 판단요소를 확정하고 이러한 각각의 요소에 대해 랜덤함수 형태로 가중치를 부여한 후, 이를 정해진 방정식을 이용하여 반복된 계산을 하는 형태로 이루어진다. 재범위험성 판단 알고리즘 작동방식에 대한 자세한 소개는 양종모, 법 분야 위험의 예측을 위한 인공지능 알고리즘과 관련된 법적 쟁점, IT와 法연구 제22집, 2021, 379-381면.

25) 주현경, 정채연, 범죄예측 및 형사사법절차에서 알고리즘 편향성 문제와 인공지능의 활용을 위한 규범 설계, 2020, 제27권 제1호, 120면.

26) 지도학습은 입력값과 함께 결과값을 같이 주고 학습시키는 방법으로 분류나 회귀 등의 분석에 주로 사용된다. 비지도 학습은 결과값 없이 입력값만을 이용하여 학습시키는 방법으로 사람들 간의 네트워크를 분석하는 작업에서 주로 사용된다. 마지막으로 강화학습은 결과값 대신에 어떤 일을 잘했을 때 보상을 주는 방식으로 학습을 시킨다. 어떤 행동을 취할 때마다 외부환경에서 보상이 주어지고 이러한 보상을 최대화 하는 방법으로 학습이 이루어진다. 이에 대한 자세한 소개는 김의중, 알고리즘으로 배우는 인공지능, 머신러닝, 딥러닝 입문, 위키북스, 2016 참조.

27) 형사법에서 인공지능의 다양한 활용가능성과 문제점을 소개한 문헌들에 따르면, 종종 머신러닝을 일반적으로 딥러닝과 같은 개념으로 소개하고 있지만, 엄밀히 말하면 딥러닝은 머신러닝과 완전히 동일한 개념은 아니다. 인공지능의 중요한 구현방법 가운데 하나가 머신러닝이고, 머신러닝 중 하나의 방법론이 딥러닝이다. 한편 딥러닝은 인공신

은 알고리즘을 사용하여 데이터를 구문 분석하고 해당 데이터에서 학습하며, 학습한 내용에 따라 정보에 근거한 결정을 내리는데, 그 기능이 무엇이든 아주 점진적으로 향상되고, 인공지능 알고리즘이 부정확한 예측을 하는 경우 설계자가 개입하여 조정해야 한다. 이에 반해 딥러닝 모델은 인간의 두뇌를 본떠서 만든 인공 신경망개념(artificial neural network)을 토대로, 알고리즘을 3계층[28]으로 구성하여 자체적으로 배우고 지능적인 결정을 내릴 수 있다. 따라서 알고리즘이 자체 신경망을 통해 예측의 정확성 여부를 스스로 판단할 수 있으므로, 설계자의 개입이 필요 없게 된다.[29] 이러한 딥러닝은 재범위험성판단에 있어서도, 인공지능이 계산기나 컴퓨터처럼 통계의 결과물을 용이하게 도출하게 해주는 등의 방식으로 단지 인간의 판단을 보조하는데 그치지 않고, 적극적으로 인간의 두뇌처럼 재범위험성 판단기준을 창출해내고 이를 발전시킬 도구로 활동할 수 있는 여지를 만들어 내고 있다.

다. 인공지능을 이용한 현실의 재범위험성 예측프로그램

인공지능을 이용한 대상자의 재범위험성 예측도구는 아직 국내에서는 본격적으로 시행되고 있지 않다.[30] 더 나아가 세계적으로도 이를 사용하고 있는 나라는 아직은 많이 찾아보기 힘든 상황이다.[31] 그러나 미국에서는 이미 이와 관련하여 인공지능 재범위험성 평가도구인 컴퍼스(Correctional Offender Management Profiling for Alternative Sanctions; COMPAS)라는 프로그램이 만들어져 많은 주들이 이런 프로그램의 도입에 대해 심각하게 고려하고 있으며,[32] 이미 몇몇 주에서는 양형과정에서 이러한 예측도구를 사용하고 있다.[33] 컴퍼스는 1989년 콜로라도 대학의 통계학 교수인 Tim Brennan과 미시건 주 Traverse City에서 교정 프로그램을 운영하던 Dave Wells에 의해 설립된 Equivant(구Northpointe)라는 회사가 개발 제공하는 재범위험성평가도구이다. 컴퍼스는 피고인의 과거 형사기록과 인터뷰를 바탕으로 미리 작성된 137개의 질문에 대한 대답을 작성한다. 137개의 질문은 과거범죄전력, 가족관계, 재정상황, 교우관계, 마약전과, 교육정도, 분노와 슬픔의 정도, 범죄에 대한 평소 생각 등 다양한 영역에 대한 피고인의 상태 및 가치관을 알아보는 문항으로 구성되어 있으며,[34] 이에 대한 대답들을 분석하여 당사자의 재범위험성을 1부터 10까지의 점수로 평가하고 있다. 판단 결과 재범위험성은 1-4점은 낮음, 5-7점은 중간, 8-10점은 높음으로 평가된다. 문제는 이러한 컴퍼스의 알고리즘이 어떻게 구성되어 있는지는 기업비밀로

경망이란 개념과 혼동하여 사용되기도 하는데, 이와 관련해서도 엄밀히 말하면 딥러닝은 인공신경망의 한 종류일 뿐이다. 이를 수식화 해보면, 인공지능⊃머신러닝⊃인공신경망⊃딥러닝으로 표현할 수 있다.

28) 3층 계층구조는 입력층(Input layer)과 중간은닉층(Hidden layer), 출력층(Output layer)으로 구성되어 있다. 입력층에서 중간은닉층으로의 연결강도와 임계치는 무작위로 추출된 수로 정해지고, 중간은닉층에서 출력층으로의 연결강도와 임계치는 학습에 의해 변경된다. 이에 대한 자세한 설명은 Tomohiro Odaka, 김성재(역), 만들면서 배우는 기계학습: 생각을 만드는 빅데이터 기술, 한빛미디어, 2012, 174면 이하 참조.

29) 이러한 딥 러닝 모델이 잘못된 결론을 도출하지 않도록 보장하는 것은, 현재의 기술로도 매우 까다로은 것으로 평가된다. 이러한 학습의 프로세스를 정확하게 만들려면 매우 많은 교육이 필요한데, 이를 보여주는 좋은 예가 바로 구글의 알파고이다. 이세돌과의 바둑대전으로 유명한 알파고는 처음에는 그 결과가 좋지 못했지만, 엄청난 양의 기보를 교육하게 되면서, 최적의 수를 이끌어 내는 정확성이 급속도로 높아졌다.

30) 우리나라 형사사법 영역에서 인공지능기술을 활용하는 수준은 아직은 걸음마 단계로 예방적 경찰활동 및 전자발찌 대상자의 효율적 관리영역 등에서, 주로 단순한 기능을 담당하는 용도로 사용되고 있다. https://m.lawtimes.co.kr/Content/Article?serial=171058

31) 유럽에서도 인공지능을 이용한 재범예측도구는 아직 공판과정에서 활용되고 있지 않다.

32) Douglas A. Berman, Are Costs a Unique (and Uniquely Problematic) Kind of Sentencing Data?, 24 FED. SENT'G REP. 159, 160 (2012) ("In some form, nearly every state in the nation has adopted, or at least been seriously considering how to incorporate, evidence-based research. . . into their sentencing policies and practices.").

33) *See, e.g.*, KY. REV. STAT. ANN. §533.010(2) (Lexis Nexis 2014); TENN. CODE ANN. §41-1-412(b) (2014); VT. STAT. ANN. tit. 28, §204a(b)(1) (2009); WASH. REV. CODE §9.94A.500(1) (2016); *cf.* ALA. CODE §12-25-33(6) (2012); 42 PA. STAT. AND CONS. STAT. ANN. §2154.5(a)(6) (West Supp. 2016).

34) https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

외부에 공개되지 않는다는 점이다. 따라서 제3자는 컴퍼스의 위험요소들이 어떻게 선정되고, 각 요소들에 얼마만큼의 가중치가 부여되는 등에 대해 알 수 없을 뿐만 아니라, 머신러닝 및 딥러닝과 관련해서도 컴퍼스가 어느 정도 발전된 알고리즘을 기반으로 하고 있으며, 과연 스스로 학습능력을 가지고 정확성을 높이고 있는지에 대해서도 구체적으로 판단하기 어려운 상황이다.[35)]

4. 비교분석

가. 비교의 전제

설득력 있는 비교분석이 이루어지기 위해서는 일단 구체적인 비교대상이 존재해야 한다. 그런데 현재 사용되고 있는 재범위험성평가도구는 K-SORAS, PCL-R등으로 특정할 수 있는 반면, 인공지능을 이용한 재범위험성 평가도구는 아직 우리나라에 존재하지 않는다. 세계적으로 눈을 돌려보아도 미국의 컴퍼스 정도가 그나마 알려진 모델인데, 이의 구체적인 알고리즘에 대해서는 앞서 살펴보았듯이 많은 정보가 비밀로 되어 있다. 또한 인공지능의 알고리즘 기술은 앞으로 시간이 지남에 따라 계속 발전될 여지가 충분하기 때문에 어떤 정도의 수준을 예상하여 비교대상으로 삼아야 할지에 대해서도 어려움이 있다. 이하에서는 이러한 고민을 바탕으로 일단 현재 미국에서 사용되고 있는 컴퍼스의 능력과 인공지능이 현재 다른 영역에서 실행되고 있는 수준의 딥러닝(예를 들어 바둑의 알파고)의 학습을 한다고 가정하였을 때, 예상되는 대략적인 모습을 기준으로 비교분석을 시도해보고자 한다.

나. 평가방식

현재 보안처분의 선고과정에서 사용되고 있는 재범위험성도구는 기본적으로 통계적 방법을 기본베이스로 하고 있다. 인공지능 재범위험성판단도구도 기본적으로 통계를 기반으로 한다는 점에서 기존의 재범위험성 평가도구와 동일하다. 물론 인공지능재범위험성판단도구는 막대한 빅데이터를 기반으로 기존보다 훨씬 많은 문항을 판단기준으로 삼을 수 있다는 점에서는, 통계적 측면에서 기존의 모델보다 더 발전된 형태로 볼 수 있을 것이다. 양자간에 본질적인 차이가 존재하는 부분은, 최종적인 판단 및 보완을 인간이 하는지 아니면 인공지능이 하는지 여부이다. 현재의 방식은 통계적 방식에 기인하여 만들어진 도구의 질문 사항에 대한 결과분석을 관련분야에 대한 전문적 식견을 갖춘 임상전문가가 판단하고 있다. 이들은 분석과정에 필요한 전반적인 계획을 세우고, 대상자에 따라 대면면담, 주변인 조사 등의 질적 방법에 의한 임상평가도 동시에 수행한다. 이 과정에서 인간이 주체가 되기에 아무래도 직관이나 선입견 등의 주관적 요소가 완전히 배제된다고 보기는 어렵다. 이에 반해 인공지능은 처음에 설계된 알고리즘을 바탕으로 철저히 기계적이고 객관적인 평가가 이루어진다.

다. 효율성

인공지능재범위험성 예측도구는 처리속도 및 비용 면에서는 현재의 재범위험성예측도구에 비해 장점이 있을 것으로 판단된다. 인공지능은 인간처럼 감정이 없고, 아프지도 않으며, 24시간 운영되기 때문에, 사건처리속도측면에서는 비교가 불가능할 정도의 압도적인 효율성을 자랑할 것으로 보인다. 이러한 인공지능재범예측도구의 장점은 특히 매우 많은 보안처분 선고사건이 발생하고, 그 수준도 단순한 설문조사에 의한 방식이 아닌, 심층적인 질적 탐구방식이 요구되는 상황에서는 더욱 두드러질 것이다. 또한 인공지능재범예

35) 이러한 문제점은 위스콘신 주의 State v. Loomis, 881 N. W. 2d 749(Wis. 2016)판결에서도 제기되었다. 이와 관련해 에이브러햄슨(Shirley S. Abrahamson)재판관은 보충의견을 통해 양측 변호사 모두 COMPAS가 어떻게 구체적으로 작동하는지에 대한 질문에 거의 대답하지 못한 점을 강조하면서, 재판부가 COMPAS의 개발 역사, 작동원리, 정확성, 유효성에 대해 매우 자세하게 설명을 요하는 노스포인트사의 변론취지서를 강제하지 않았던 점은 잘못이라고 판단했다.

측도구는 비용적인 측면에서도 장점이 있을 것으로 예상된다. 현재 보안처분의 선고과정에서 참여하고 있는 평가 전문가의 숫자나 임금 및 인공지능 개발비용에 대한 구체적인 자료를 구하지 못한 관계로 비용편익에 대한 자세한 비교는 불가능하다. 하지만 일반적인 경험법칙으로 생각해보면 평균임금액과 관련사건수는 높아지거나 큰 변화가 없는 반면, 시간이 지날수록 인공지능 기술은 보편화되고 발전되므로 이를 활용할 수 있는 단가는 낮아질 것으로 예측된다. 따라서 인공지능재범위험성도구를 사용하는 것이 비용절감 효과도 클 것이고, 이는 한정된 예산을 바탕으로 비용문제에 점점 민감해지는 사법현실에서 중요한 고려요소가 될 것으로 판단된다.

라. 정확성

사실상 가장 중요한 요소인 재범예측의 정확성부분은 결론적으로 말하면 현재로서는 일률적으로 평가하기 어렵다고 판단된다. 우선 이에 대해서는 경험적인 연구결과가 뒷받침되어야 한다. 한국의 재범위험성평가도구들의 타당성에 대한 연구는 지속적으로 이루어지고 있다. 대부분의 방식은 각 문항의 항목이 재범위험성 판단에 실질적인 영향을 미치는지와 판단의 기준점으로 삼은 점수[36]가 재범위험성 판단기준으로서 유효한지 등을 로지스틱 회귀분석을 통해 판단하고 있다. 이에 따르면 연구에 따라 특정기한에 한꺼번에 출소한 범죄자를 대상으로 한 것이 아니라는 점이나, 관찰기간도 길지 않았다는 등의 한계가 있긴 하지만,[37] 현재 사용되고 있는 재범위험성평가도구는 전반적으로 약 70%정도의 정확성을 보이고 있다.[38] 예를 들어 성폭력범죄자에 대한 재범 위험성도구인 K-SORAS에 대한 타당성 연구결과를 보면, 529명의 성범죄자들을 대상으로 특정기간 재범 여부를 분석한 결과 이들 전체의 성범죄 재범률은 10.6%이었고, 항목별 점수합산 13점 이상의 고위험군으로 평가받은 사람들의 재범률이 12점 이하의 점수를 받은 사람들보다 유의미한 수준에서 높게 나타났다.[39] 또한 13점 이상의 점수를 받은 사람들의 실제 재범률은 72.2%정도로 나타났다.[40] 동일한 연구진이 행한 2차 연구에서는 163명의 성범죄자를 대상으로 재범여부를 추적하였는데, K-SORAS 15점 이상의 고위험군 가운데 재범을 저지른 확률은 72.1%였다.[41]

문제는 인공지능재범위험성평가도구에 대한 경험적 연구결과이다. 우리나라는 아직 이를 시행조차 하지 않고 있기 때문에, 당연히 이에 대한 경험적 연구결과가 존재하지 않는다. 따라서 이 부분 역시 현재는 미국의 컴퍼스를 중심으로 살펴볼 수밖에 없다. 이에 대한 연구결과에 따르면 컴퍼스의 판단이 법률 비전문가 인간의 판단보다 결코 더 정확한 것이 아니라고 보고되고 있다. 대표적인 것은 뉴햄프셔주 다트머스 대학교의 Julia Dressel와 Hany Farid가 2018년에 Science Advances지에 소개한 연구로,[42] 이 연구는 플로리다주의 브로워드 카운티(Broward County)의 범죄데이터를 기반으로 컴퍼스와 일반사람들의 판단을 비교하였다.[43] 해당연구는 2013년과 2014년에 브로워드 카운티에 수감되었던 7214명의 범죄자들의 데이

36) 예를 들어 K-SORAS는 피검사자의 연령, 혼인관계, 본 범죄의 특성, 전과 특성에 대한 총 15개의 항목에 대해 각각 점수를 매겨 0점부터 29점까지 총점을 산출한 후, 점수에 따라 재범위험성 수준을 3등급으로 분류하고 29점 만점에서 13점 이상인자는 고위험자로 간주한다.

37) 고려진, 성범죄자들에 대한 재범추적연구, KORAS평가를 중심으로, 168면 이하.

38) 성폭력 범죄를 대상으로 적용되는 재범위험성 평가도구에 대한 국외의 연구결과에 대한 요약적 소개는 이수정、황의갑、박선영, 재범방지를 위한 범죄자처우의 과학화에 관한 연구(Ⅰ) - 위험성평가도구 개발에 관한 연구, 45-62면 참조.

39) 고려진, 성범죄자들에 대한 재범추적연구, KORAS평가를 중심으로, 155-170면.

40) 고려진, 성범죄자들에 대한 재범추적연구, KORAS평가를 중심으로167면.

41) 이수정, 고려진, 최혜림, 재범추적을 통한 한국성범죄자위험성평가척도(KSORAS) 타당도 연구, 999-1016면. 이 연구에서 로지스틱 회귀분석을 통한 AUC(Area Under the ROC curve)는 .676이였다. AUC지수는 항상 0과 1사이로 1.0은 오류 없이 완벽하게 예측했다는 것을 의미한다. 일반적으로 AUC지수가 .70이상이면 비교적 우수한 예측평가로 인정받는다고 한다.

42) Julia Dressel and Hany Farid, The accuracy, fairness, and limits of predicting recidivism, 2018 SCIENCE ADVANCES.

터 가운데 1000명의 데이터를 무작위로 추출하여 이들이 석방 후 2년 내에 다시 범죄를 저질렀는지를 추측하는 형태로 이루어졌다.[44] 이 연구에 참여한 사람들은 법률전문가가 아닌 사람들로 구성되었다.[45] 실험은 총 두 번 이루어 졌는데, 1차 실험은 실험데이터의 인종 요소를 고려하지 않고 이루어졌고, 2차 실험은 인종요소를 고려하여 이루어졌다. 판단 방법과 관련하여서는 컴퍼스는 앞서 살펴본 것처럼 137개의 질문에 대해 대답을 한 결과를 종합하여 판단을 한 반면, 인간 참여자들은 실험 데이터 범죄자들의 성별과 나이 그리고 과거범죄경력의 정보만을 토대로 판단하게 하였다.

[표1] 1000명 대상자에 대한 사람과 컴퍼스 알고리즘 예측 비교, 전체 정확도는 퍼센트로 표기, 다양한 임계값에서 모델의 분류성분 측정(AUC-ROC), 민감도 범위(d'), 편향(β)[46]

	(A) 사람 (인종요소 고려 없음)	(B) 사람 (인종요소 고려)	(C) 컴퍼스
정확도 (전체)	67.0 %	66.5 %	65.2 %
AUC-ROC (전체)	0.71	0.71	0.70
d'/β(전체)	0.86/1.02	0.83/1.03	0.77/1.08
정확도 (흑인)	68.2 %	66.2 %	64.9 %
정확도 (백인)	67.6 %	67.6 %	65.7 %
거짓 양성 (흑인)	37.1 %	40.0 %	40.4 %
거짓 양성 (백인)	27.2 %	26.2 %	25.4 %
거짓 음성 (흑인)	29.2 %	30.1 %	30.9 %
거짓 음성 (백인)	40.3 %	42.1 %	47.9 %

위의 [표1]을 살펴보면, 재범 확률 예측의 정확도에 있어서 인간의 판단이 컴퍼스의 판단보다 오히려 미세하나마 더 정확한 것임을 알 수 있다. 인간은 1차 실험에서는 67%, 2차 실험에서는 66.5% 예측 정확도를 보였는데, 이에 반해 컴퍼스의 재범예측정확도는 65.2%였다. 또한 실험 표본을 흑인범죄자와 백인범죄자로 나누어 보더라도 인간참가자는 흑인 범죄자와 백인 범죄자에 대해 각각 68.2%와 67.6%의 정확성을 보인 반면, 컴퍼스는 64.9%와 65.7%를 나타내었다. 더 세부적으로 재범을 저지를 것으로 판단했는데 실제로는 저지르지 않은 경우를 나타내는 거짓양성(false positive)과 그 반대를 나타내는 거짓 음성(false negative)에서도 컴퍼스의 판단이 인간의 판단보다 전반적으로 뛰어나다고 볼 수 없는 것으로 나타났다. 이 밖에도 위의 표에는 나오지 않지만 인간이 20명씩 그룹을 만들어 집단적으로 토의하여 판단한 결과도 67%로 컴퍼스의 평균 재범예측정확도인 65.2%보다 높았다.[47] 이렇게 컴퍼스에 한정해서 비교해보면 인공지능 재범예측도구가 인간의 직관보다도 그 정확성면에서 뛰어나지 않으며 대략 70%의 정확성을 나타내는 우리나라 재범위험성예측도구와 비교해도 우수하지 않다는 결론이 도출된다. 그러나 인공지능은 계속 발전 중에 있기 때문에, 훨씬 많은 빅데이터를 기반으로 보다 구체적으로 알고리즘을 설정하거나, 인공지능프로그램 도입 이후에도 딥러닝을 통해 지속적으로 오류에 대해 학습할 경우 정확도가 상승할 여지는

43) 본 연구는 총 두 번의 실험으로 이루어 졌는데, 1차 실험은 피의자의 인종을 고려하지 않고 이루어졌고, 2차 실험은 피의자의 인종을 고려하여 이루어졌다.
44) 1000명의 데이터는 다시 각50명씩 총 20개의 그룹으로 만들어 참여자들에게 분배되었다.
45) 1차 실험에서는 참여자 462명 중 400명의 답변이 유효답변으로 연구에 활용되었고, 2차 실험에서는 449명의 참여자 중 400 명의 답변이 유효하게 연구에 활용되었다.
46) Julia Dressel and Hany Farid, The accuracy, fairness, and limits of predicting recidivism, 2018 SCIENCE ADVANCES. https://www.science.org/doi/10.1126/sciadv.aao5580 2면 Table 1 발췌번역.
47) 이에 대해서는 Julia Dressel and Hany Farid, The accuracy, fairness, and limits of predicting recidivism, 2018 SCIENCE ADVANCES. 2면.

충분히 존재한다. 이 경우는 현재 시행되고 있는 재범위험성평가와 비교해서도 더 우수한 정확성을 나타낼 가능성도 있다고 생각된다. 왜냐하면 지금의 재범위험성평가는 기본적으로 통계를 바탕으로 이루어지는데, 통계적인 측면에서 인공지능은 빅데이터를 토대로 훨씬 더 정확하고 관련성이 높은 결과를 도출할 수 있기 때문이다. 특히 인공지능재범예측도구의 정확성과 관련해서 생각해 보아야 할 부분은 처음 알고리즘을 설계하는 설계자의 경험과 능력이 중요하게 반영될 것이라는 점이다. 결국 알고리즘도 인간이 기존의 전문지식을 바탕으로 설계하는 것이기 때문에, 기본적으로는 현재재범위험성 도구에서 사용되고 있는 문항이나 지식을 바탕으로 판단요소가 설계될 수밖에 없다. 따라서 기술적인 부분과 전문지식에 풍부한 설계자들이 함께 모여 지속적으로 개발을 시도한다면 분명 현재의 재범위험평가보다 정확성이 높은 평가도구를 만들 수 있다고 생각된다.[48] 다만 이 경우에도 보안처분에서 판단되는 재범위험성평가의 정확도를 획기적(예를 들어 90%이상)으로 높이기는 어렵다고 판단된다. 왜냐하면 보안처분에서의 재범위험성평가는 대상자의 특정행동(재범)을 예측하는 것으로 그 자체도 매우 변수가 많은데, 판단기간도 특정 기간(예를 들어 6개월내)이 아닌 대상자의 남은 전 기간 동안을 대상으로 하므로 고려요소가 너무 많기 때문이다.[49] 또한 각 범죄와 관련하여 수집 가능한 데이터양에 따라서도 정확성은 한계가 있을 수 있다. 알고리즘의 예측 능력은 의미 있는 상관관계(예를 들어 과거에 A라는 행동을 한 사람은 미래에도 A라는 행동을 한다.)를 도출할 수 있는 양질의 훈련 데이터에 매우 영향을 받는다. 따라서 정확성을 높이기 위해서는 매우 많은 데이터를 필요로 한다. 하지만 테러범죄와 같은 특정범죄 같은 경우는 그 수가 상대적으로 적어 재범위험성을 예측하는 데에 한계가 있을 수 있다.

정확성의 비교 문제에서 마지막으로 고려되어야 하는 점은 인공지능은 통계에 근거한 패턴분석에 의존하다보니 자연스럽게 집단적 특성을 우선 고려하게 되어 현재의 재범위험성평가에 비해 개인적인 특성파악이 상대적으로 소홀히 될 수 있다는 점이다. 이러한 개인적 특성들은 숙련된 전문가가 수치화된 요소 이외에도 대면면담이나 관찰을 통해 파악될 여지도 있는 만큼, 이러한 점에서도 인공지능은 약점을 보일 수 있다. 따라서 현재 재범위험성도구 방법 가운데 개별면담을 통한 임상적 판단 부분은 인공지능재범위험성예측도구가 도입되더라도 이를 보완하는 방식으로 계속 사용되는 것이 타당할 것으로 판단된다.

5. 소결.

인공지능을 이용한 재범위험성평가는 현재 사람이 하고 있는 재범위험성평가와 비교해 그 작동방식의 특징으로 인해 훨씬 방대한 양의 통계처리와 객관적인 처리가 가능하다. 따라서 인공지능 재범예측도구는 처리속도 및 비용적인 면에서는 현재의 재범위험성예측도구에 비해 커다란 장점이 있어 도입의 필요성(실익)이 있다고 판단된다. 그러나 재범위험성의 정확도 측면에서는 어느 정도의 정확성 향상을 기준으로 하는지

48) 판단요소는 컴퍼스의 설문문항에서도 볼 수 있듯이 범죄 전력, 학력, 재산, 최초 범행 연령 등으로 구성될 수 있는데, 이러한 판단요소들의 수가 인공지능재범예측도구가 현재의 재범예측도구보다 훨씬 많다. 이는 단순히 몇 개의 요소만 사용하는 경우보다 시간이 지나서 데이터가 쌓이면 정확성면에서 장점을 보일 수 있다. 하지만 이와 관련해서도 섬세한 접근이 필요한데, 왜냐하면 이러한 판단요소의 수를 계속해서 늘린다고 해서 계속해서 예측력이 정비례하여 높아진다는 보장이 없기 때문이다. 결국 지속적인 결과값에 대한 통계분석을 통해 판단요소를 몇 개까지 고려하는 것이 최적의 값인지를 찾는 연구와 다양한 판단 요소들 가운데 어떤 요소가 재범위험성 판단의 정확성을 높이는데에 결정적인 역할을 하는지를 찾는 연구가 인공지능의 분석 및 학습기능을 통해 지속적으로 이루어져야 한다.

49) 이는 현재 재범위험성평가에서도 마찬가지로 내포되어 있는 문제이다. 정신과 의사로서 전문감정으로 활동하고 있는 독일의 Nobert Nedopil에 의하면 예측의 시간적 흐름과 관련하여 유의미한 변화결과를 도출할 수 있는 시간은 대략 6개월 정도라고 한다. 즉, 6개월 정도가 흐르면 그 사이에 피관찰자의 재범위험성에 대한 유의미한 변화가 목격될 수도 있다는 것이다. Peter Asprion, Gefährliche Freiheit?: Das Ende der Sicherungsverwahrung, Herder, 2013, 94면. 이러한 특징은 보안처분의 가해제, 예를 들어 우리나라의 경우, 전자발찌 부착이후 임시해제 신청에 따른 검사를 얼마의 시간간격을 두고 해야 하는지와 관련해서도 중요한 참고사항이 될 수 있다.

에 따라 도입의 실익이 있을지 여부가 결정될 것으로 판단된다. 현재 미국에서 사용되고 있는 컴퍼스를 기준으로 할 때에는 현재의 재범예측도구에 비해 결코 더 우수하다고 볼 수 없다. 물론 시간이 지나면 특정 조건의 충족(빅데이터의 수집량 상승, 인공지능의 분석능력 및 딥러닝을 통한 자가 학습능력 향상, 설계자의 재범위험성관련 전문지식과 알고리즘 설정능력 향상)으로 인해 현재의 수준보다는 높은 재범예측의 정확성을 나타낼 수 있으므로, 장래를 기준으로는 도입의 실익이 있다고 판단된다.[50] 그러나 아무리 인공지능이 발전한다고 하더라도 특정기한의 정함이 없이 사람이 특정범죄를 저지를 개연성을 평가한다는 것은, 너무 많은 고려요소로 인해 본질적으로 많은 오류가능성을 내포하고 있다. 따라서 애초부터 인공지능재범예측도구의 정확성에 대한 기대를 무오류에 가깝게 설정한다면, 이는 지나친 기대가 될 것이다.

III. 인공지능재범위험성예측도구의 규범적 문제점에 대한 검토

1. 인공지능재범위험성예측도구의 도입 전 쟁점

인공지능을 통한 재범위험평가 실제로 보안처분의 선고과정에서 사용되기 위해서는 법관이 인공지능을 통해 도출된 재범위험성결과를 공판과정에서 판단의 근거로 사용될 수 있는지의 문제부터 검토되어야 한다. 만약 이것이 불가능하다면, 인공지능을 통한 재범예측 결과의 정확성이 아무리 우수하다고 하더라도 해당 자료는 의미가 없기 때문이다.

가. 증거능력

우리형사법은 증거재판주의를 천명하고 있다. 형사소송법 제307조에 따르면 범죄사실은 반드시 증거능력이 있는 그리고 법률이 정한 적법한 증거조사절차를 거친 증거에 의하여 인정되어야 하며, 이는 합리적인 의심이 없는 정도의 증명에 이르러야 한다고 하여 증명력이 있는 증거에 의한 사실인정을 통한 엄격한 증거재판주의를 요구하고 있다. 이와 관련해 만약 수사기관에서 유무죄의 판단증거로 알고리즘의 예측결과를 제출할 경우, 이에 대한 증거능력이 인정될 것인지와 인정된다면 어떠한 요건 하에서 어느 정도의 증명력을 부여할 수 있는지에 관하여 많은 논의가 이루어지고 있다.[51] 왜냐하면 인공지능 알고리즘의 예측결과는 이른바 '과학적 증거(scientific evidence)'에 해당한다고 할 수 있는데, 이러한 과학적 증거의 증거능력을 인정할 수 있는지에 대해 형사소송법상 명확한 기준이 없기 때문이다.[52] 그런데 이러한 논의를 유무죄 판단이 아닌 보안처분(예를 들어 특정 성범죄자에 대해 위치추적 전자장치의 부착명령)을 부과할 것인지를 판단함에 있어서도 적용하여, 인공지능재범위험성평가결과에 대한 증거능력을 인정할 수 있을지에 대한 문제제기가 있다.[53] 일단 보안처분의 선고는 유무죄 판단의 영역이 아닌 것은 확실하다. 보안처분은 유죄를 전제로 이루어지고, 유죄의 결과로 형벌을 부과하면서 형벌의 예방적 흠결을 위해 함께 부과되는 제2의 형사제재이다. 따라서 크게 보면 양형의 영역이라고도 볼 수 있다. 양형 판단을 위하여 제출되는 자료

---

50) 현재 재범위험성도구 방법 가운데 개별면담을 통한 임상적 판단 부분은 인공지능이 도입되더라도 인공지능재범위험성예측도구를 보완하는 방식으로 여전히 사용되는 것이 타당할 것이다.

51) 대표적으로는 김웅재, 형사절차에서 인공지능 알고리즘의 활용가능성과 그 한계, 서울대학교 기술과 법센터, LAW & TECHNOLOGY, 제16권 제4호, 2020년 7월, 17면 이하.

52) 판례는 거짓말 탐지기의 경우 증거능력을 부정하나, 유전자 감식결과 등이 문제된 사건들에서는 증거능력과 증명력을 엄격하게 구분하지 않은 채 어떠한 조건 하에서 강한 증명력을 부여할 수 있는지에 초점을 맞추는 태도를 보이고 있다. 거짓말 탐지기에 대해서는 대법원 1986. 11. 25. 선고, 85도 2208 판결, 대법원 2005. 5. 26. 선고, 2005도130 판결 등 참조; 과학적 방법을 이용한 유전자 감식에 대해서는 대법원 2007. 5. 10. 2007도 1950판결, 대법원 2011. 5. 26선고, 2011도1902판결 참조.

53) 김웅재, 형사절차에서 인공지능 알고리즘의 활용가능성과 그 한계, 서울대학교 기술과 법센터, LAW & TECHNOLOGY, 제16권 제4호, 2020년 7월, 20면이하.

에 대해서는 형사소송법상 엄격한 증명의 법리[54]가 적용되는지, 즉 그 자료가 증거능력을 갖추어야 하는지에 관하여는 견해대립이 있다.[55] 판례는 이와 관련해"법원은 범죄의 구성요건이나 법률상 규정된 형의 가중·감면의 사유가 되는 경우를 제외하고는 법률이 규정한 증거로서의 자격이나 증거조사방식에 구애됨이 없이 상당한 방법으로 조사하여 양형의 조건이 되는 사항을 인정할 수 있다."고 하여 양형에 관한 사실은 엄격한 증명의 대상이 아니라는 입장을 취하고 있다. 따라서 이에 따를 경우 양형에 관한 자료는 반드시 형사소송법상 증거능력을 구비할 필요가 없고 법원이 상당하다고 인정되는 방법으로 자료를 수집 조사할 수 있으므로, 인공지능에 의한 재범위험성판단결과도 자유롭게 사용할 수 있는 결과가 도출된다.[56]

그러나 이렇게 인공지능에 의한 재범위험성판단결과를 자유롭게 보안처분 공판과정에서 사용할 수 있다는 결론에는 동의하지만, 처음의 문제제기 및 결론에 이르는 논리전개과정은 타당하지 않다고 판단된다. 왜냐하면 보안처분에서의 인공지능재범위험성판단결과는 양형사실인지와 상관없이 애초에 증거능력문제와는 다른 영역의 문제로 보는 것이 타당하기 때문이다. 형사소송법에서 증거(證據)라 함은, 기본적으로 사실관계를 확실하게 하기 위하여 사용되는 자료를 말한다.[57] 그런데 보안처분에서 판단되는 재범위험성판단결과는 엄밀히 말하면 사실관계를 확정짓는 자료가 아니라 일종의 지표[58]로서 법관의 최종적인 가치판단(보안처분 부과여부)을 돕기 위한 의견서이다. 재범위험성이 낮다거나 높다는 것은 어떤 사실에 대한 평가이지 그 자체가 사실은 아니기 때문이다. 물론 재범위험성평가도 구체적으로 보면, 당사자가 과거에 어떤 특정범죄를 범했다는 부분적인 사실이 나오지만, 이 사실은 그 자체를 증명하고자 하는 것이 아닌, 궁극적으로 당사자의 재범위험성이 낮다거나 높다는 것을 판단하기 위한 하나의 자료로 사용된 것이므로 재범위험성판단결과가 사실관계를 확정짓는 자료라고 보기는 어렵다. 따라서 결론적으로 보안처분과 관련된 공판과정에서 인공지능재범위험성판단결과를 수사기관에서 제출하는 것은 증거로 보기 어렵고, 이에 따라 증거능력을 통한 사용제한 문제는 애초에 논의할 필요가 없다고 보는 것이 타당하다.

나. 공정한 재판을 받을 권리

증거능력과는 상관이 없지만, 이러한 사실만으로 인공지능재범위험성평가도구의 결과가 공판과정에서 자유롭게 사용될 수 있는 것은 아니다. 왜냐하면 이와 관련해서는 피고인의 절차적 참여권 및 방어권이 문제가 될 수 있기 때문이다. 헌법 제27조 제1항은 "모든 국민은 헌법과 법률이 정한 법관에 의하여 법률에 의한 재판을 받을 권리를 가진다."라고 규정함으로써 모든 국민에게 적법하고 공정한 재판을 받을 권리를 보장하고 있다. 헌법재판소는 이러한 공정한 재판을 받을 권리를 해석함에 있어, "신속하고 공개된 법정의 법관의 면전에서 모든 증거자료가 조사·진술되고 이에 대하여 피고인이 공격·방어할 수 있는 기회가 보장되는 재판, 즉 원칙적으로 당사자주의와 구두변론주의가 보장되어 당사자가 공소사실에 대한 답변과 입증 및 반증을 하는 등 공격, 방어권이 충분히 보장되는 재판을 받을 권리"라고 설시하고 있다.[59] 이에 따르면 보안처분도 재판과정을 통해 선고되므로, 이 과정에서 피고인에게 절차적 참여권과 방어권이 보장되어야

54) '엄격한 증명'은 법률상 증거능력이 있고 적법한 증거조사를 거친 증거를 통해 확인하는 증명방식이다. 이에 반해 '자유로운 증명'이란 증거능력이나 적법한 증거조사를 요하지 않는 증거에 의한 증명을 말한다. 자유로운 증명의 경우에는 증거능력을 요하지 않을 뿐만 아니라 증거조사의 방법도 법원의 재량에 맡겨져 있기에 법률에 규정된 증거조사절차에 의하지 않고 변론종결 후에 접수된 서류나 전화 등을 통해 확인된 증거에 의하여도 사실을 인정할 수 있게 된다. 배종대, 홍영기 형사소송법 제2판, 홍문사 305면.

55)견해대립의 구체적 내용에 관하여는 신동운, 신형사소송법, 법문사 1021면.

56) 이와 같은 주장으로는 김웅재, 형사절차에서 인공지능 알고리즘의 활용가능성과 그 한계, 서울대학교 기술과 법센터, LAW & TECHNOLOGY, 제16권 제4호, 2020년 7월, 20면 이하.

57) 형사재판이 공정하게 진행되려면 먼저 범죄에 대한 사실관계가 확정되어야 하기 때문이다.

58) 재범의 위험성평가도구는 형태적으로 증거가 아니므로, 이와 관련해 증거능력 배제나 증명력 탄핵논의는 부적절하다는 주장에 대해서는 양종모, 법 분야 위험의 예측을 위한 인공지능 알고리즘과 관련된 법적 쟁점, IT와 法연구 제22집, 2021, 396면.

59) 헌재 1996. 12. 26. 94헌바1, 판례집 8-2, 808, 819-820.

함은 이견이 있을 수 없다. 문제는 피고인이 이러한 참여권과 방어권을 행사하기 위해서는 일단 상대방의 주장과 근거에 대해 이해하고 반박 할 수 있는 가능성이 보장이 되어야 하는데, 인공지능의 특성상 과연 이것이 가능한지여부이다.[60]

인공지능은 앞서 작동원리에서도 살펴보았듯이 설계된 알고리즘을 바탕으로 움직인다. 그리고 이 알고리즘은 보통 영업비밀로 되어 있다. 실제로 컴퍼스와 관련된 분쟁에서도 컴퍼스를 운영하는 회사는 이러한 알고리즘을 법원에게 제출하지 않았다. 그러나 이러한 영업비밀의 문제는 입법을 통해 해당 알고리즘을 공판과정에서 공개할 수 있게 만들 여지가 충분히 있다. 우선 인공지능재범예측도구의 제작을 굳이 사기업이 아닌 국가차원에서 하는 방법을 생각해 볼 수 있다. 이 경우 자연스럽게 개인의 재산권 보호라는 민감한 문제를 피해갈 수 있다. 또한 만약 컴퍼스와 같은 민간기업의 프로그램을 사용하더라도 처음부터 알고리즘 공개의 동의하에만 공판과정에서 사용되도록 입법을 통해 민간기업의 참여를 유도할 수도 있다. 물론 이 경우 사기업들의 반발이 심할 수 있지만, 시장 경쟁의 측면에서 보면 이러한 부분적인 불이익을 감수하고도 다른 이익(예를 들어 회사 홍보효과 등)을 이유로 기본 알고리즘을 충분히 공개할 수도 있기 때문이다. 이렇게 된다면 일단 알고리즘의 투명성 측면에서는 정당성이 확보될 수 있게 된다. 그런데 더 중요하면서 본질적인 문제점은, 인공지능재범위험성도구를 개발하여 제공한 주체가 이런 알고리즘을 공개하고 그 과정을 설명하고 그 설명을 검증하고 싶어도 물리적으로 이것이 불가능한 경우이다. 이는 특히 인공지능의 머신러닝(딥러닝)의 특성과 관계가 있는데, 머신러닝과 같은 복잡한 알고리즘의 맥락에서는 처음에 설정된 변수가 공개가 되어도, 이후 학습을 통해 스스로 진화하기 때문에 어떻게 특정한 결과 값이 나오게 되었는지의 과정을 파악하기 어렵게 만든다.[61] 이러한 문제점을 해결하기 위해 최근 처음 알고리즘을 설계할 때부터 알고리즘이 결론 또는 예측에 도달하는 방법을 설명하는 시스템을 설계하는 시도가 이루어지고 있다. 일명 설명 가능한 인공지능모델인데, 지속적인 노력에도 불구하고 아직까지는 제대로 구현되지 못하고 있는 실정이다. 또한 이런 부분이 설사 성공한다고 하더라도, 설명가능성이 높아지면, 예측의 정확도는 떨어지는 것이 일반적이므로, 이 경우 잘못하면 처음부터 인공지능재범예측도구를 도입할 실익자체가 없어질 수 있다.[62] 따라서 이러한 점들을 고려해보면, 현재 공정한 재판을 받을 권리위배와 관련해 제기되고 있는 문제점은 충분히 설득력이 있으며, 이에 따라 현재수준의 인공지능재범예측도구는 공판과정에서의 활용불가능하다고 보는 것이 타당하다.

하지만 이 경우, 인류 역사상 거대한 진보의 흐름이라고 볼 수 있는 인공지능 알고리즘의 사용을 포기하는 것이어서 선뜻 찬동하기 어렵다는 견해도 존재한다.[63] 이는 관련기술이 매우 빠르게 발전하고 있고, 앞으로 우리 삶에서의 인공지능의 활용은 결코 막을 수 없는 흐름이라는 점을 감안해보면 전혀 설득력이 없는 것은 아니다. 사실 각 영역에서 설명 가능한 알고리즘을 만들려는 수많은 노력들도 결국 따지고 보면 인공지능은 우리 삶에서 피할 수 없는 영역이 될 것이고 효율성이나 정확성 측면에서도 기존보다 나은 결과를 가져올 것이라는 것을 확신이 있기 때문에 많은 비용에도 불구하고 계속하여 시도되고 있는 것이기 때문이다. 이런 점을 고려해보면 보안처분 영역에서 재범위험성예측도구를 활용하기 위해, 피고인의 방어권의 침해를 최대한 약화시키는 방향으로 논거를 제시하면서 이러한 인공지능재범예측도구를 계속 사용하는 방법도 시도 될 수 있다. 이는 실제로 컴퍼스와 관련하여 위스콘신주 대법원이 적법절차원칙에 위배되지 않는다고 결정을 내리면서 판시한 방법으로, ①인공지능재범예측도구는 부정확한 정보에 기초하여 재범위험성 판단이 이루어진 것이 아니라는 점 ②인공지능재범예측도구의 정확성은 이미 객관적으로 검증되었

---

60) 한애라, 사법시스템과 사법환경에서의 인공지능 이용에 관한 유럽 윤리헌장의 검토 -민사사법절차에서의 인공지능 도입 논의와 관련하여 -, 저스티스 172호, 2019. 68면.

61) 현재로서는 심지어 알고리즘을 최초에 설계한 프로그래머조차도 완전한 파악이 어려운 상황이라고 판단하는 것이 다수의 평가이다.

62) https://web.archive.org/web/20180421234120/http://www.ciokorea.com/news/34654

63) 양종모, 법 분야 위험의 예측을 위한 인공지능 알고리즘과 관련된 법적 쟁점, IT와 法연구 제22집, 2021, 398면.

다는 점 ③인공지능재범예측도구의 판단문항확인을 통해 피고인에게는 이의 제기권이 원칙적으로 부여되어 있다는 점 ④인공지능재범예측도구를 통해 나온 결과는 어디까지나 법관의 최종판단에 있어 보조적인 자료에 불과하다는 점을 그 논거로 하고 있다.[64)]

각 논거들을 조금 더 구체적으로 살펴보면, ①번은 인공지능재범예측도구의 평가요소 항목들은 피고인이 직접 기입하는 질의응답으로 되어 있는데, 이와 관련해 피고인이 직접기재를 하거나 사법기관의 공적기록 등을 확인하여 정보가 기입되기 때문에 적어도 인공지능재범도구에 처음에 설정된 각 항목의 수치 값은 정확하다는 것을 강조하는 논거이다. ②번은 이미 시행하고 있는 다른 주나 연구를 통해 해당인공지능재범예측도구의 판단요소들 및 판단기준의 유효성이 검증되었고 결과의 정확성 부분에서도 일정수준의 목적을 달성한다는 것을 확인받았다는 점을 강조함으로써 알고리즘 설계자체가 잘못 설정되어 있지 않다는 것을 주장하는 논거로 판단된다. ③번은 인공지능 재범예측도구는 최초에 설정된 판단요소들을 근거로 작동이 되는데, 그 요소들은 이미 공개되어 피고인도 알고 있기 때문에 적어도 이와 관련된 부분에 대한 이의제기는 가능하다는 점이다. 예를 들어 인공지능 재범예측도구는 가족관계를 재범예측의 하나의 요소로 삼고 있는데, 재범예측과 가족관계는 특별한 관련이 없음에도 재범예측도구는 있다는 것을 전제로 운영되고 있으므로 이를 기반으로 이루어진 결과는 신뢰할 수 없다는 주장을 피고인은 할 수 있다는 것이다. 마지막으로 ④번은 설사 이러한 재범예측도구의 결과가 도출되었고, 이러한 도출과정에 대한 명확한 설명은 불가능하다고 하더라도, 이 결과는 피고인에 대한 보안처분의 선고에 직접적인 근거가 아니라 하나의 자료에 불과하다는 논리이다. 따라서 애초에 이 자료자체가 보안처분 선고에 큰 영향을 미치지 않으므로 그 자료결과에 대한 설명이 조금 부족한 점이 있어도, 그것이 적법절차원칙(공정한 재판을 받을 권리)에 위배될 정도는 아니라는 것이다. 이러한 논거들을 투명성, 설명가능성, 검증가능성 문제를 해결하는 본질적인 논거라고 보기는 어렵다. 왜냐하면 위의 4가지 논리를 모두 인정하다고 하더라도, 결국 최종 결과에 이르는 과정을 모두 설명할 수 없다는 점은 여전히 명확하고, 보안처분의 선고에 인공지능재범예측도구의 결과가 직접 영향을 미치지는 않았다고 하더라도, 법관의 심증형성에 일정한 영향을 끼칠 수 있는 여지가 존재하기 때문이다. 더구나 다른 논거를 통해 인공지능재범예측도구의 정확성은 어느 정도 객관적으로 검증되었다고 판단하였으므로, 이는 법원이 해당조사결과를 완전히 무시하였다고 보기 어려운 논거로 사용될 수도 있다. 따라서 위의 항변을 통해 피고인의 공정한 재판을 받을 권리가 침해되는 문제를 완전히 해결할 수는 없으므로, 이는 결국 보안처분을 선고받아야 하는 당사자의 기본권 보호 이익과 인공지능의 사용을 통해 얻을 이익을 비교형량하여 판단할 문제로 생각된다. 이와 관련해 현재의 인공지능재범예측도구 수준으로는 그 정확성이 사람의 재범예측방법과 비교해 유의미하게 높다고 볼 수 없으므로, 보안처분 선고과정에서 인공지능재범예측도구를 도입하는 것은 헌법상 보장된 공정한 재판을 받을 권리를 침해한다고 판단된다.

3. 인공지능재범위험성평가도구 도입 후 쟁점.

가. 인공지능 재범위험성 평가결과의 의무적인 반영문제

만약 장래에 인공지능기술이 비약적으로 발전해 설명 가능한 인공지능 모델이 도입된다고 하면, 이를 통한 결과물을 어떻게 활용할 것인지의 문제가 제기 될 수 있다. 이와 관련해 우선 생각해 볼 수 있는 규범적 쟁점은, 법관이나 검사가 보안처분을 선고하거나 청구함에 있어서 이러한 인공지능재범예측도구를 통해 당사자의 재범위험성을 판단하는 것을 강제규정으로 설정해야 하는지의 문제이다. 참고로 현재 우리나라의 보안처분은 이에 대한 강제규정을 두고 있지 않다. 예를 들어 전자발찌의 경우, 전자장치 부착 등에 관한 법률 제6조 제1항에 따르면 “검사는 부착명령을 청구하기 위하여 필요하다고 인정하는 때에는 피의자의

64) State v. Loomis, 881 N. W. 2d 749(Wis. 2016), Ann Walsh Bradley 재판관 법정의견.

주거지 또는 소속 검찰청 소재지를 관할하는 보호관찰소의 장에게 범죄의 동기, 피해자와의 관계, 심리상태, 재범의 위험성 등 피의자에 관하여 필요한 사항의 조사를 요청할 수 있다."고 규정되어 있고, 제6조 제5항은 "검사는 부착명령을 청구함에 있어서 필요한 경우에는 피의자에 대한 정신감정이나 그밖에 전문가의 진단 등의 결과를 참고하여야 한다."고 하여 필요한 경우에만 감정인의 재범위험성평가를 참고하도록 하고 있다. 이런 필요한 경우는 검사 스스로 판단하는 것이므로, 결국 감정인의 판단은 강제 규정이 아닌 재량규정으로 해석된다. 판사와 관련해서는 아예 '필요한 경우에는'이라는 규정자체도 없다. 동 법률 제9조 1항은 "법원은 부착명령 청구가 이유 있다고 인정하는 때에는 다음 각 호에 따른 기간의 범위 내에서 부착기간을 정하여 판결로 부착명령을 선고하여야 한다."하여 검사가 전자발찌명령을 청구함에 감정인의 재범위험성 평가를 첨부하지 않더라도, 판사 또한 이를 판단하지 않고 보안처분을 선고할 수 있도록 규정하고 있다.[65]

문제는 법관과 검사는 법률의 전문가이지만, 재범위험성 판단의 전문가는 아니라는 점이다. 보안처분은 일반적으로 대상자가 책임원칙에 따라 선고된 형벌의 집행을 종료하였음에도 불구하고 사회 안전예방을 위해 오직 특정범죄에 대한 재범위험성을 근거로만 부과되는 형사제재이다. 이러한 보안처분은 대상자에 대한 기본권 침해가 매우 강하기 때문에 최후 수단적으로만 부과되어야 한다. 이러한 최후수단의 범위를 설정하는 것은 결국 재범위험성예측의 정확성이다. 그리고 이런 예측을 인공지능시대에서는 인공지능재범위험성예측도구가 하는 것이다. 앞서 살펴보았듯이 인공지능재범위험성예측도구는 빅데이터와 현재의 전문지식을 바탕으로 한 알고리즘 설계능력에 의해 현재의 재범예측도구에 비해 시간이 지나면 더 정확한 판단을 할 개연성이 높다. 그렇다면 이 경우는 관련 법률을 개정하여 법관 혹은 검사가 이러한 재범위험성평가를 의무적으로 참고하도록 하는 것이 타당하다고 생각한다.[66] 이와 관련해 물론 현재의 법률에 따르더라도 법관이나 검사가 필요에 따라 이러한 재범위험성판단검토를 할 수 있으므로 이는 큰 실익이 없다는 반론이 제기될 수도 있다. 그러나 할 수 있는 것과 해야 하는 것은 차이가 있다. 의무 규정으로 설정해 놓으면 각 사건마다 빠짐없이 판단을 해야 하지만, 재량규정은 그렇지 않을 구멍을 만들어 놓는 것이기 때문이다. 보안처분을 선고받는 당사자의 입장에서 보더라도 인공지능재범예측도구의 설명가능성 부분이 해결이 된다면, 법관이 자신의 직관보다 구체적이고 논증여지가 있는 판단결과를 참고하는 것이 유리할 것이다.[67] 또한 사건의 처리측면에서 인공지능은 뛰어난 효율성을 자랑하므로, 각 사건마다 재범위험성 평가가 의무규정으로 되어 있어도 큰 행정적 문제는 발생하지 않을 것으로 예측된다.

나. 법관의 인공지능 재범위험성평가도구에 대한 구속여부

위의 문제와 연결되어 더 나아가 법관이 이러한 인공지능에 의한 재범위험성판단결과에 구속될 필요는

---

65) 다른 보안처분들도 전부 비슷하게 규정되어 있다.

66) 참고로 독일의 경우 이미 현재의 재범위험성평가방법에 있어서도 보호감호나 치료감호의 선고에 있어서는 감정인의 판단을 의무규정으로 두고 있다. 그 취지는 재범위험성판단의 정확도를 높이고 보호감호를 선고받는 대상자의 기본권을 더욱 중요하게 보호하기 위함이라고 한다. 관련조문 독일형사소송법§246a,§414 Abs. 3,§415 Abs. 5.

67) 형사소송법 제308조에 따라 법관은 자유심증주의에 따라 보안처분의 선고유무를 판단할 자유를 가진다. 그러나 이는 그것이 실체적 진실발견에 적합하기 때문이지 법관의 자의적인 판단을 인용한다는 것은 아니다. 판례도 이와 관련해 법원의 합리적 판단과 객관적 논증의무를 인정하고 있다."비록 사실의 인정이 사실심의 전권이더라도 범죄사실이 인정되는지는 논리와 경험법칙에 따라야 하고, 충분한 증명력이 있는 증거를 합리적 이유 없이 배척하거나 반대로 객관적인 사실에 명백히 반하는 증거를 근거 없이 채택·사용하는 것은 자유심증주의의 한계를 벗어나는 것으로서 법률 위반에 해당한다. 또한 범죄의 유무 등을 판단하기 위한 논리적 논증을 하는 데 반드시 필요한 사항에 대한 심리를 다하지도 아니한 채 합리적 의심이 없는 증명의 정도에 이르렀는지에 대한 판단에 섣불리 나아가는 것 역시 실체적 진실발견과 적정한 재판이 이루어지도록 하려는 형사소송법의 근본이념에 배치되는 것으로서 위법하다."대법원 2016. 10. 13., 선고, 2015도17869, 판결.

없는지에 대한 문제도 제기될 수 있다. 이러한 물음이 제기되는 배경에는 앞서 살펴본 것처럼 법관은 법률판단의 전문가이지, 재범위험성판단의 전문가가 아니라는 점이 놓여 있다. 재범의 정확성과 이에 따른 보안처분의 정당성을 더욱더 확보하기 위해서라도, 재범위험성 판단에 있어서는 인공지능의 재범위험성평가 결과에 구속되는 것이 타당하지 않겠냐는 논지인 것이다.[68] 그러나 장래에 인공지능이 발전하여 재범위험성 판단에 있어 지금보다 훨씬 더 전문성을 가지더라도 법관이 이에 구속되는 것은 타당하지 않다고 생각된다. 왜냐하면 재범위험성 판단과 보안처분의 선고는 헌법이 규정한 법관의 특수한 지위, 즉 법관의 독립성하고도 관련되며, 이는 다시 시민의 법관에게 재판받을 권리와 연결되기 때문이다. 재범위험판단의 전문성 측면에서는 법관보다 인공지능이 더 우위에 있을 수 있다. 그러나 인공지능재범위험성예측도구에 최종적인 결론을 맡길 경우 다음의 문제들이 발생할 수 있다. 첫째, 설계자의 의도에 따라 인공지능재범위험성 도구의 결과가 달라질 가능성이 존재한다. 예컨대 대상자에 대해 같은 판단요소(예를 들어 나이, 전과 사실 등)를 초기 값으로 설정한다고 하더라도 설계자가 처음에 알고리즘을 계획할 때, 어떤 가치를 더 우위에 두는지에 따라 인공지능끼리도 결과가 달라질 수 있다.[69] 또한 빅 데이터의 선정과 판단요소의 선정, 각 판단요소들에 대한 가중치 설정 등과 관련해서도 인공지능마다 견해가 다를 수 있는 가능성이 존재한다. 이 경우 최종적인 판단기준을 설정할 수 없게 된다. 둘째, 설사 다양한 인공지능재범예측도구의 결과가 합치된다고 하더라도 항상 외부의 압력에 휘둘릴 가능성이 존재한다. 예를 들어 재범위험성예측인공지능의 설계자나 회사끼리 특정이익을 위해 담합을 하여 다 같이 프로그램 알고리즘을 특정집단에 유리하게 설정하는 경우가 있을 수 있다. 이렇게 항상 재판에서 가장 중요한 공정성이 훼손될 가능성이 존재하기에, 우리 헌법은 이를 방지하고자 헌법조문[70]에서 직접 법관의 독립성과 지위를 보장하여 이 문제를 해결하고 있다. 물론 법관도 사람이기에 공정성의 시비가 늘 문제될 수 있지만, 이 경우도 제척, 기피, 회피제도 등을 통해 최대한 시스템적으로 재판의 공정성을 확보하려고 노력하고 있다. 따라서 이러한 현재체계와 미래에 인공지능이 최종판단을 맡게 될 경우를 비교해 본다면, 재범위험성평가에 대한 인공지능의 판단을 법관이 참고는 반드시 하되, 여기에 구속은 되지 않는 형태로 법률내용이 규정되는 것이 타당하다고 생각된다.[71]

VI. 결론

이상의 논의를 요약하면 다음과 같다.

보안처분의 판단에 있어 인공지능재범위험성예측도구는 현재의 재범위험성예측도구와 비교해 처리속도 및 비용적인 측면에서는 우위에 있는 것으로 판단된다. 중요한 재범위험성예측의 정확성 측면에 있어서는 현재 개발된 수준이 미국의 컴퍼스를 기준으로 할 경우, 현재의 우리나라에서 사용되고 있는 재범위험성예측도구에 비해 그 기능이 뛰어나다고 보기 어렵다. 다만 인공지능 기술은 워낙에 빨리 발전하고 있고, 수집된 빅 데이터의 양도 시간이 갈수록 늘어나며, 재범위험성 평가도구를 설계하는 인간의 관련지식도 시간이 갈수록 발전하기 때문에 인공지능에 의한 재범위험성평가의 정확성은 지속적으로 상승할 것으로 예측된다. 하지만 아무리 인공지능이 발전한다고 하더라도 매우 장기간을 대상으로 인간의 행동을 예측해야 하는 재

68) 물론 이는 장래에 기술발달로 인해 인공지능의 재범위험성예측의 정확성과 설명가능성부분이 개선되었다는 것을 전제로 하는 것이다.

69) 예를 들어 A인공지능은 재범위험성 판단과정에서 상대적으로 안전을 더 추구하는 입장에서 접근하도록 설계되어 있을 수도 있고, B인공지능은 반대로 개인의 자유를 더 추구하는 입장으로 설계되어 있을 수 있다.

70) 헌법 제103조: 법관은 헌법과 법률에 의하여 그 양심에 따라 독립하여 심판한다. 헌법 제106조 제1항: 법관은 탄핵 또는 금고 이상의 형의 선고에 의하지 아니하고는 파면되지 아니하며, 징계처분에 의하지 아니하고는 정직ㆍ감봉 기타 불리한 처분을 받지 아니한다.

71) 현재 법률에 의해서도 법관은 감정인의 재범위험성 판단결과에 구속되지는 않는다. 다만 현재 법률은 법관이 강제적으로 참고를 하도록 규정되어 있지는 않다.

범위험성예측의 특성상 일정 수준이상으로 비약적으로 재범위험성 정확도가 올라 갈 것을 기대하기는 힘들 것이다. 이렇게 미래에는 효율성과 정확성 측면에서 인공지능재범위험성예측도구가 현재의 재범위험성예측도구에 비해 장점을 가질 것이기 때문에 인공지능재범예측도구의 도입필요성은 긍정된다. 다만 중요한 것은 인공지능재범예측도구가 도입이 될 경우 발생할 수 있는 규범적 문제점들이다. 이와 관련하여 첫 번째 쟁점은 유무죄 판결에서와 같이, 보안처분의 선고에 있어서도 인공지능을 통해 판단된 재범위험성 결과를 법관이 판단의 근거로 사용하기 위해서 증거능력이 필요한지의 문제였다. 이와 관련해 보안처분의 재범위험성판단을 통해 도출된 자료는 사실관계의 확정을 위한 자료가 아니므로 애초에 증거능력이 문제될 필요가 없다고 보는 것이 타당하다. 따라서 법관이 공판과정에서 인공지능에 의한 재범위험성판단결과를 자유롭게 고려하여 판단을 내릴 수 있다. 두 번째는 인공지능이 가진 알고리즘의 불투명성과 설명 및 검증 불가능성으로 인해 당사자의 공정한 재판을 받을 권리가 침해되는지 여부이다. 이와 관련해 위스콘신주대법원은 4가지 근거를 통해 권리침해를 부정했지만, 인공지능의 머신러닝(딥러닝)이 가지는 복잡성과 불투명성으로 인해, 설명 가능한 인공지능 모델이 개발되기 전에는 공정한 재판을 받을 권리가 침해된다고 보는 것이 타당하다. 따라서 미국의 컴퍼스 수준의 인공지능 재범위험성도구를 현재 우리 형사법에서 사용할 수는 없다. 그러나 미래에 인공지능 기술이 발전되어 설명가능성 문제가 해결된다면, 인공지능 재범위험성예측도구는 보안처분 선고과정에서 사용될 여지가 있다. 이 경우 인공지능재범위험성 평가결과를 법관이 의무적으로 참고해야 하는지, 그리고 더 나아가 이 결과에 구속되어야 하는지가 문제 될 수 있다. 보안처분은 형벌 못지않게 당사자에 대한 기본권 침해가 크고, 주로 형벌의 집행을 마친 자를 대상으로 선고되는 제재인 만큼 최후 수단적으로 사용되고, 당사자의 기본권 보호에 충실해야 한다. 이런 점을 고려해본다면, 법관의 경험적 직관에만 의존하는 것보다는 보다 더 높은 재범위험성예측의 정확성을 보여주는 인공지능재범위험성판단결과를 의무적으로 참고하는 것이 타당하다 할 것이다. 그렇지만 법관이 이 결과에 구속될 필요까지는 없다고 생각된다. 그것은 우리 헌법이 재판의 공정성을 위해 시스템적으로 보장하고 있는 법관의 특수한 지위와 독립성의 장점이 훼손되는 것이기 때문이다. 따라서 인공지능재범위험성판단결과를 법관이 반드시 참고는 하되, 그 결과에 구속되지는 않는 방향으로의 법률개정이 요구된다.

< 토 론 문 >

인공지능재범위험성평가도구의 도입필요성과 규범적 문제점

박노섭(한림대학교 글로벌학부)

발제자는 사회전반적인 흐름과 인공지능을 비롯한 첨단 과학기술에 대한 투자비중과 국민적 호응도를 보면 앞으로 인공지능이 형사정책에서 차지하는 비중은 다른 영역에서처럼 점점 높아져 갈 것으로 평가하고, 인공지능을 통한 재범예측가능성과 예측결과에 대한 규범적 문제를 검토하고 있다는 점에서 시사하는 바가 매우 크다. 발제자는 본 논문에서 인공지능을 통한 재범위험성 평가와 관련하여 먼저, 인공지능을 통한 재범예측의 객관성 및 정확성 제고에 관한 관점을 분석하고, 그 다음 판단과정의 투명성과 설명가능성 등의 새로운 규범적인 관점에서도 처분 영역에서 인공지능의 활용가능성과 한계에 대해 검토하고 있다.

토론자는 이러한 발제순서에 따라 토론하고자 한다.

먼저, 발제자가 서술한 바대로 인공지능을 이용한 재범예측 분야는 기본적으로 통계를 바탕으로 한 예측 프로그램이 중심이 되고 있는 만큼 빅데이터를 기반으로 하는 인공지능이 결합되기 용이하다. 그러나 발제자가 본문 내용에서 설명하고 있듯이 현재의 인공지능 수준(사람의 판단보다 낮은 수준)에 미루어 보아 인공지능에 의한 재범평가결과가 재범위험성 평가도구를 이용한 평가결과를 대체하기는 신뢰도나 타당도가 낮아 당장 실현될 가능성은 낮아 보인다.

이와 관련하여 발제자는

“ 딥러닝은 재범위험성판단에 있어서도, 인공지능이 계산기나 컴퓨터처럼 통계의 결과물을 용이하게 도출하게 해주는 등의 방식으로 단지 인간의 판단을 보조하는데 그치지 않고, 적극적으로 인간의 두뇌처럼 재범위험성 판단기준을 창출해내고 이를 발전시킬 도구로 활동할 수 있는 여지를 만들어 내고 있다.” 라고

서술하고 있어 추후 관련 데이터가 많아지고, 데이터를 활용한 인공지능 모델이 개발될 경우 이를 극복할 수 있을 것으로 판단하고 있다.

또한, 발제자는

“인공지능재범위험성판단도구는 막대한 빅데이터를 기반으로 기존보다 훨씬 많은 문항을 판단기준으로 삼을 수 있다는 점에서는, 통계적 측면에서 기존의 모델보다 더 발전된 형태로 볼 수 있

*을 것이다."*라고 설명하고 있다.

재범위험성 평가도구에 의한 평가의 경우 평가전문가가 제시된 기준에 적합한 척도에 따른 점수를 본인의 경험을 바탕으로 부여하는 방식(검사자간 일치도 평가를 통한 신뢰도 측정)이다. 이러한 평가방식을 인공지능에 의한 평가로 대체하기 위해서는 척도에 따라 점수를 부여한 결과인 데이터가 머신러닝이 가능할 정도로 상당한 량이 필요하나, 평가 특성상 만만한 작업이 아니다. 보안처분의 영역에서 인공지능에 의한 재범위험성판단도구가 형사실무에 본격적으로 활용되기 위해서는 선결과제로 관련 데이터를 어떻게 확보할 것인가에 대한 구체적인 방안에 대한 연구(학습데이터 구축방안)가 선행될 필요가 있어 보인다. 더구나, 심층적인 질적탐구의 방식으로 진행되는 재범평가와 같은 영역에서 정제된 데이터 확보하고 가공하는 작업이 결코 쉬운일이 아니다.

뿐만 아니라 인공지능 재범위험성판단도구의 정확성은 학습데이터의 정확도와 매우 밀접하다. 그러나 본문 내용에 의하면 재범위험성 평가도구에 의한 평가의 신뢰도가 72%의 정도에 불과하다.[1)]

관련하여 발제자는

*"인공지능은 계속 발전 중에 있기 때문에, 훨씬 많은 빅데이터를 기반으로 보다 구체적으로 알고리즘을 설정하거나, 인공지능프로그램 도입 이후에도 딥러닝을 통해 지속적으로 오류에 대해 학습할 경우 정확도가 상승할 여지는 충분히 존재한다. 이 경우는 현재 시행되고 있는 재범위험성 평가와 비교해서도 더 우수한 정확성을 나타낼 가능성도 있다고 생각된다."*라고 설명하고 있다.

그러나 인공지능의 경우 점수가 수치로 부여된 데이터를 학습한 결과를 통한 예측이라는 점을 감안하면 이 경우에 발생하는 정확도 문제를 어떻게 개선할 것인가도 핵심문제로 판단된다. 추가적인 설명이 필요하다.

두 번째 분야인 규범적인 문제와 관련하여 토론자는 보안처분에 대한 최종적인 판단자는 해당사건을 담당하고 있는 판사라는 점에 주목하면,

*"현재 공정한 재판을 받을 권리위배와 관련해 제기되고 있는 문제점은 충분히 설득력이 있으며, 이에 따라 현재수준의 인공지능재범예측도구는 공판과정에서의 활용불가능하다고 보는 것이 타당하다."*라는 발제자의 의견에 전적으로 공감하는 바이다.

재범위험성 평가도구에 의한 평가결과 *경우* 통계를 기반으로 하는 분석기법이다. 따라서 **①데이터-②규칙-③해답의 과정을 거치게 되고,** 해답에 이르게 되는 과정이 복잡하다고 하더라도 전문

1) 발표문에 의하면 한 연구에서 163명의 성범죄자를 대상으로 재범여부를 추적하였는데, K-SORAS 15점 이상의 고위험군 가운데 재범을 저지른 확률은 72.1%였다.

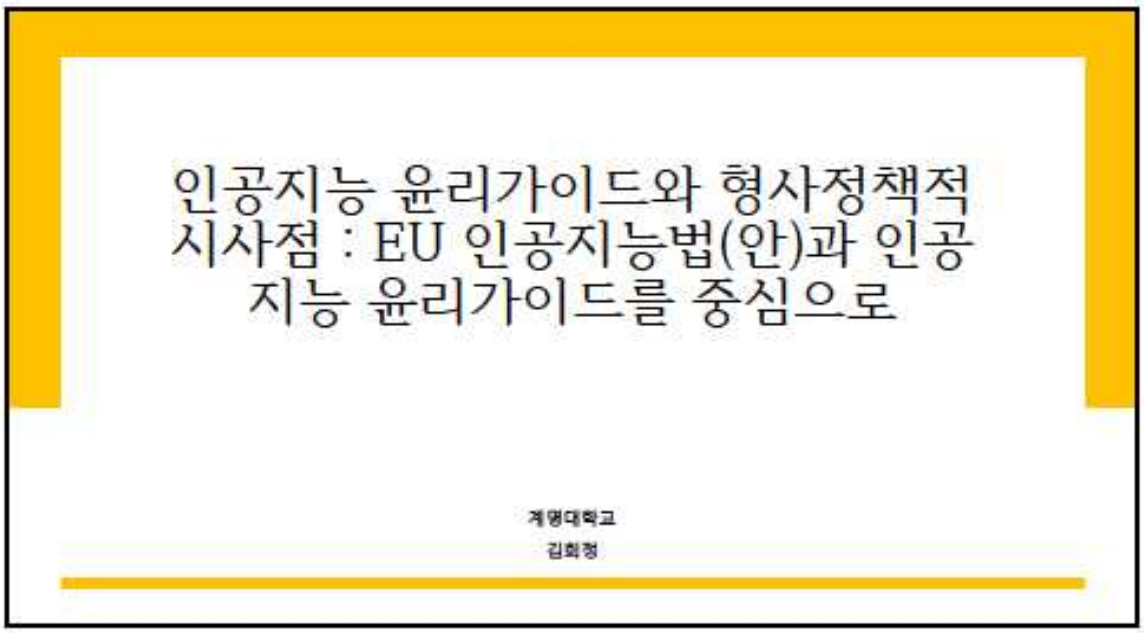
가인 감정인의 도움을 받으면 그 결론에 이르는 과정을 이해할 수 있고, 이를 판단의 기초로 활용하는 것은 물론 가능하다.
그러나 머신러닝의 경우 **①데이터-③해답-②규칙의 과정을 거치면서** 머신 러닝은 데이터가 주어지면, 기계가 스스로 데이터로부터 규칙성을 찾는 것에 집중하는 핵심적인 업무(Task)이다. 일단 규칙성을 발견하고 나면, 그 후에 들어오는 새로운 데이터에 대해서 발견한 규칙성을 기준으로 정답을 예측하게 되는데, 이러한 방식은 기존 규칙기반의 프로그래밍 방식으로 접근하기 어려웠던 문제에 대한 새로운 해결책을 제시해 준다. 그러나 어떤 과정을 거쳐 해당 규칙이 만들어졌는가를 설명할 수 없는 경우가 대부분이다. 더구나, 머신러닝의 하나인 딥러닝의 특징은 비지도 학습(Unsupervised Learning)을 통한 데이터 분류 방식이다. 지도 학습의 경우 자료수집자인 사람이 이미 그나마 해답을 알고 있어서 최종 결과인 규칙(패턴)을 나름 예측할 수 있다. 그러나 비지도 학습 경우 어떤 해답을 제시함이 없이 정보를 입력하고 컴퓨터가 알아서 패턴(분류)을 찾아내는 방식이라는 점에서 최종적인 결론에 이르는 과정을 설명하기는 더욱 어렵다. 예를 들어 바둑게임에서 알파고가 어느 지점에 바둑돌을 왜 놓았는지를 분석해내기는 사실상 불가능에 가깝다.

앞에서 말한 대로 형사재판에서 사실판단이든 보안처분이든 최종적인 판단자인 판사의 입장에서는 합리적인 의심이 없을 정도로 해당 사안에 대한 이해를 요구한다. 이런 점에서 본다면 인공지능에 의한 재범위험성 평가결과는 향후 정확도가 현저히 높아진다고 가정을 하더라도 단독으로 활용은 불가능할 것으로 보인다. 다만, 법관이 다른 자료나 증거에 근거하여 판단을 한 후 - 의료계에서 핫이슈가 되었던 IBM의 왓슨(WATSON)처럼 - 이에 대한 검증을 위한 보조수단으로 활용은 가능해 보인다.

< 발 표 자 료 >

인공지능 윤리가이드와 형사정책적 시사점
: EU 인공지능법(안)과 인공지능 윤리가이드를 중심으로

계명대학교 김희정



Ⅰ. 인공지능의 개념

인공지능, 머신 러닝, 딥러닝?

- 인공지능이란?

사고나 학습 등 인간이 가진 지적 능력을 컴퓨터를 통해 구현하는 기술
특정한 목적 달성을 위해 환경 분석을 수행하고 어느정도 자율성을 기반으로 지능적 행동을 수행하는

- 머신 러닝이란?

컴퓨터가 스스로 학습하여 인공지능의 성능을 향상시키는 기술 방법

- 딥 러닝이란?

인간의 능력과 비슷한 인공신경망 방식으로 정보를 처리하는 것

알고리즘의 개념과 특성

(1) 알고리즘이란?

- 알고리즘이란 특정한 문제를 풀기 위해 거쳐야 하는 일련의 과정들을 묶어 놓은 절차나 방법. 즉, 머신 러닝을 통해 AI가 학습한 규칙을 알고리즘이라고 하며, 인공지능은 알고리즘을 익히며 스스로 판단과 유추를 하게 되는 것임

(2) 알고리즘의 특성

권력성: 알고리즘이 인간의 문화, 가치관을 바꿀 수도 있으며, 인간을 판단하여 중요한 권리를 침해할 수 도, 지대한 영향을 미칠 수 도 있음

정치성: 알고리즘을 통해 인간의 가치관과 생각, 세계관 등이 통제될 수 있고, 이런 통제를 통해 알고리즘이 하나의 정치적 도구로 이용 될 수 있기에 정치성을 가지고 있다고 할 수 있음

상업성과 생활침투성: 많은 전자상거래에서 알고리즘 광고,소비자 가격 담합 등으로 알고리즘이 영리목적으로 이용되고 있고, 알고리즘이 인간의 취향과 기호를 분석하여 음악, 영화, 뉴스 등 다양한 서비스나 상품 등을 추천하고 인간은 이를 부담 없이 받아들이는 편이기에 알고리즘은 상업성과 생활침투성의 특성이 있다고 할 수 있음

비밀성: 알고리즘이 영업상 비밀 등의 이유로 어떻게 만들어 졌는지 비밀로 하는 경우가 많아 비밀성의 특성을 가지고 있다고 볼 수 있음

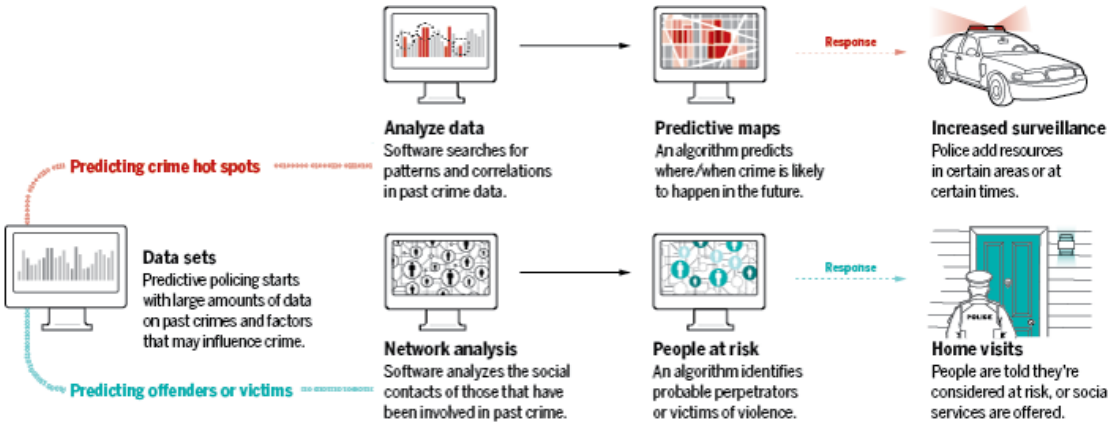
Ⅱ.형사사법에서의 인공지능 활용현황

제주경찰청, 신변보호용 안면인식 CCTV 전국 첫 시범 운영

▲ 제주경찰청이 전국에서 처음으로 시범 운영 중인 신변보호용 인공지능 CCTV의 작동 단계. 제주

인공지능과 프레드폴(PredPol)

프레드폴은 인공지능(AI) 이 범죄유형, 발생한 위치, 날짜, 시간 등 과거 범죄가 발생한 데이터를 AI로 분석해 미래에 발생할 범죄유형과 지역을 예측함



국내 프리카스

- 프리카스는 그동안 축적된 112 신고, 설치된 폐쇄회로(CC)TV 위치, 교통사고 건수, 유흥시설 숫자, 학교, 공원 등은 물론 인구, 기상, 실업률, 고용률 등 각종 데이터를 자동으로 분석해 범죄 위험도를 예측하는 시스템으로, 특정 요일, 특정 시간에 어디에서 범죄가 발생할 확률이 높은 지 분석해 실시간으로 알려줌
- JG는 과거 범죄가 수차례 일어나 3등급 내 고위험군 위험지역이라는 뜻이고, TS는 지역 주민이 순찰을 요구한 위치를 말함.
- 화살표를 클릭하면 그곳에서 발생했던 과거 범죄·사건 등을 구체적으로 확인할 수 있음
- 프리카스가 연동된 업무용 휴대전화에 담당 경찰관이 이상 유무를 체크하면 해당 정보는 실시간으로 업데이트됨

[경찰, 과학과 만나다]① "여기서 우회전하세요"... 우범지역 찾아주는 AI, 순찰 풍경 바꿨다 - 조선비즈 (chosun.com)

법무부 인공지능 전자감독시스템

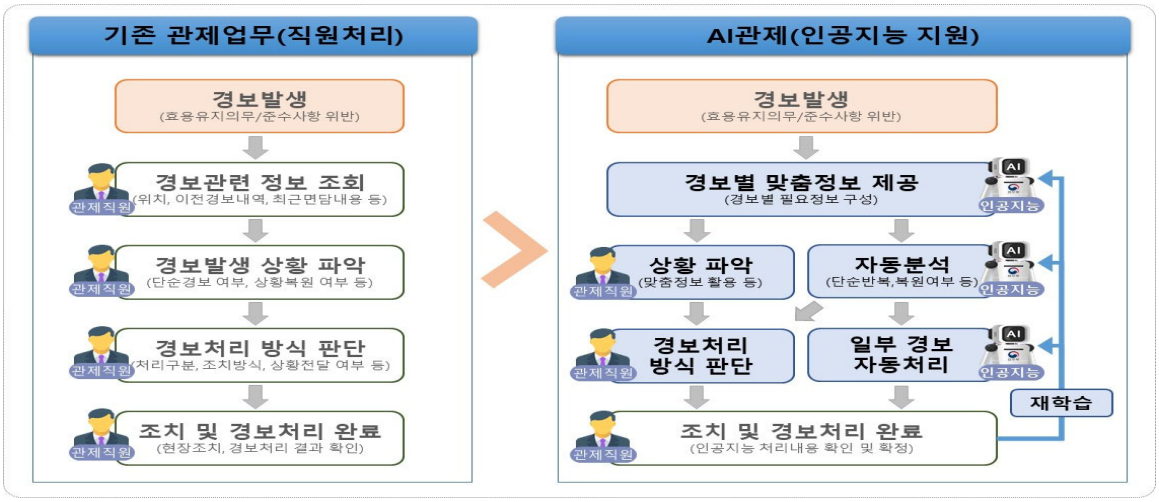
- 법무부는 증가하는 전자감독 대상자의 효율적 관리와 재범 방지의 선제적 개입 역량을 높이기 위해 빅데이터와 인공지능(AI) 기술을 활용한 전자감독시스템을 도입·운영

  ✓ 전자감독제도는 '08년 성폭력범죄에 처음 시행된 후 '09년 미성년자 유괴범죄, '10년 살인범죄, '14년 강도 범죄, '20년 가석방되는 모든 사범으로 확대되고 전자보석제도까지 도입되어, 연중 집행사건이 '08년 205건에서 '21년 5월 7,373건으로 33배 이상증가

  ✓ 구속 기소된 피고인에 대해 전자장치 부착을 조건으로 보석을 허가해주는 전자장치 부착 조건부 보석제도 시행('20.8.)

  ✓ 한편 고위험 성범죄자 1:1전담보호관찰제도입, 야간 미귀가 지도 등 늘어나는 업무에 비해 전담직원 인력의 부족한 상황은 계속되고 있어, 전자감독 업무의 효율적 운영을 지원하기 위해 인공지능·빅 데이터 분석을 활용한 첨단기술을도입

인공지능(AI) 기반의 전자감독서비스



출처: 법무부

인공지능 보호관찰

- 인공지능 보호관찰은 전자감독 대상자의 범죄정보, 이동패턴, 면담내용 등의 데이터와 유흥가 지역 등공간.지리정보를 분석하여 대상자 주요 특징 및 최신 동향,재범위험성 평가 등의 정보를 종합적으로 제공해줌
- 전자감독 대상자의 주요 특징 및 최근 동향을 보다 신속하게 파악할 수 있게 되어 특이 대상자 집중관리 등 선제적 감독 역량이 강화됨
- 인공지능 보호관찰에서 제공되는 기능

✓ 대상자 측위 밀도분석에 의한 낮과 밤, 주중과 주말의 이동패턴

✓ 공간 지역속성 분석을 통한 유흥가 등 주요 체류 지역 정보

✓ 조사서와 면담내용 토대로 대상자 주요특성, 면담이슈 자동추출

✓ 이전 경보발생 패턴, 범죄수법분석표 제공, 재범위험성 평가요인 해설, 업무 안내(톡) 지원 등

* 보호관찰을 받을 사람으로는 형의 선고유예를 받은 사람(형법 제59조의2), 형의 집행유예를 선고받은 사람(형법 제62조의2), 가석방되거나 임시 퇴원한 사람(형법 제73조의2, 보호관찰 등에 관한 법률 제22조 · 제23조), 보호처분을 받은 사람(소년법 제32조1항4호 · 5호), 다른 법률에서 보호관찰을 받도록 규정된 사람

인공지능 보호관찰

출처: 법무부

Ⅲ. 형사사법에서의 인공지능 윤리 문제

1. 편향성

재범예측평가 프로그램 뿐 아니라 범죄 예측에서도 편향성이 문제가 됨

경찰이 특정 커뮤니티에 대해 차별적 관행을 하여 특정 지역을 '고위험군'으로 불필요하게 또는 잘못 분류할 수 있음

이는 과거 데이터 세트의 편향성은 인공지능 학습에 상당한 영향을 미치게 되고, 결과적으로 범죄 예측 치안분야에서의 인공지능 이용은 편향된 분석을 악화시킬 수 있음

데이터의 오류

- AI 알고리즘은 예측 치안 유지에 사용되어 범죄 활동에 대한 대량의 과거 데이터를 식별하고 분류하여 위험한 사람이나 장소를 판단함.
- 이러한 프로세스는 위험 평가 또는 위협 평가라고도 하는데 일반적으로 이런 인공지능의 활용 의도는 좋지만 이러한 알고리즘을 제공하는 과거 데이터가 상당한 우려를 불러일으킬 수 있음
- 데이터 입력자가 데이터를 잘못 입력할 수도 있고, 범죄 데이터 자체가 범죄 전체를 담을 수 없는 다소 부분적 정보이며, 신뢰하기 어려운 정보도 상당부분 있기 때문에 문제가 될 수 있음
- 특정지역이나 인종 등에 대해 데이터가 편향적일 수 있다는 문제는 끊임없이 제기되어 왔음

형사사법에서의 인공지능의 편향성

- COMPAS 알고리즘이 유죄판결을 받은 범죄자가 재범하는 특정한 경향을예측하는 AI
- COMPAS를 통해, 흑인 범죄자들은 백인 범죄자들보다 거의 두 배나 더 높은 위험으로 분류되지만 실제로는 반대의 결과가나타남
- 미국의 판사들은 COMPAS의 재범위험 평가를 토대로 보석금 책정 뿐 아니라 형량과 관련된 것들을 결정하면서 피고인의 재범가능성 등을 결정

Figure 2–5: 'COMPAS Software Results', Julia Angwin et al. (2016)

2. 안면인식기술과 윤리 문제

- **보안상의 문제(Data Security)**

✓ 사람의 얼굴은 다른 개인정보와 달리 암호화 하거나 변경이 어려워 사생활 침해, 스토킹, 신원도용 등의 문제가 발생할 가능성이 높음

✓ 상업적, 공익목적 안면인식기술에 대한 해킹가능성은 언제나 있으며, 특히 정부 기관 등에서 사용하는 공익목적의 안면인식기술은 보통 단일 공급업체에서 만들어지기 때문에 하나만 보안상의 문제가 생기면 다수의 피해가 발생될 위험이 있음

- **투명성 결여(Lack of Transparency)**

✓ 안면인식은 지문 인식과 달리 다른 생체 인식과 달리 원격으로 비밀리에 쉽게 캡처 될 수 있음

- **오용(Misuse)**

✓ 편향성, 차별 문제 등

✓ 인식 오류가 높은 시스템은 오식별 우려를 야기하지만, 오류가 적은 높은 수준의 기술은 반대로 보안 감시 능력을 증가시킴

✓ 누군가를 잘못 식별한 안면인식의 오류는 앞서 설명한 바와 같이 형사사법에서 여러 문제를 발생시켰음.

2020년 1월 미시간주, 로버트 윌리엄스는 파밍 턴 힐스에 있는 그의 집 진입로에서 경찰에 체포되었는데, 디트로이트경찰은 윌리엄스가 지역 시놀라(Shinola) 매장에서 여러 개의 시계를 훔쳤다는 혐의로 그를 체포하였음. 디트로이트 경찰은 시놀라 매장의 CCTV에 찍힌 흑인 남성의 영상 자료를 미시간 주 경찰에게 보냈고, 미시간 주 경찰의 안면인식 시스템은 윌리엄이 CCTV 속의 흑인 남성일 가능성이 높다고 판단함 – 하지만 윌리엄은 강도 범죄를 저지른 자가 아닌 무고한 시민이었고 주 경찰에게 안면인식 프로그램을 판매한 데이터 웍스 플러스 회사는 윌리엄을 용의자로 지목한 방법을 설명하지 않음

3. 인공지능 기술의 기본권 침해

패턴 식별을 통한 AI 범죄예방을 위한 다양한 방안이 고안되고 있지만 이미 범죄예방을 위한 여러 형사정책에 대한 법적, 윤리적 문제가 지적되어 왔고, 이런 문제제기는 AI를 통해 점점 커질 것으로 예측됨

안면 인식 데이터와 관련된 데이터 침해는 신원 도용, 스토킹 및 괴롭힘의 가능성이 높으며, 인공지능의 투명성도 낮음

안면인식기술(FRT)를 사용하여 자신의 지식이나 동의 없이 개인을 식별하는 것은 특히 생체인식이 개인에게 고유하기 때문에 프라이버시 문제를 제기됨

- **프라이버시권 침해**

✓ 안면인식기술(FRT)는 마스크를 쓰거나 사진에서 얼굴을 흐리게 해도 여전히 개인을 식별할 수 있을 정도로 발전함

✓ 프라이버시 전문가들은 FRT를 피하는 것이 비현실적이라는 점을 감안할 때, 전방위적인 감시는 표현의 자유, 집회의 자유 등을 침해할 가능성이 높다고 지적함

✓ 콜로라도 법대 Margot Kaminski 교수는 "당신이 어디에 있든 당신의 얼굴을 추적할 수 있는 정부는 당신이 어디에 있든 당신의 위치를 추적할 수 있으며, 이는 당신이 가진 모든 연관성을 추적할 수 있다는 것을 의미한다고 설명하였음

(1) 미국

- 안면인식 알고리즘의 사용에 대해 미국시민자유단체들은 이 기술이 미국을 지나치게 억압적인 감시국가로 몰아갈지도 모른다는 우려와 함께 정부가 악용할 가능성을 제기함

(2) 한국

- 국내에도 최근 시민단체들이 법무부·과기부가 얼굴인식 인공지능식별추적 시스템 구축 사업을 추진하면서 출입국 본인확인용으로 수집·보유하고 있는 내외국인 1억 7천만 여건의 얼굴정보를 민간기업의 인공지능 학습 및 검증용으로 무단 제공하여 인권과 기본권을 침해했다고 주장하며 공익감사청구 하였음

- 이들은 얼굴 등 생체인식 정보를 활용하는 **고위험 인공지능에 대하여 모라토리엄** 등 규제를 구체화하고 있는 유럽 및 국제인권기구의 입장에도 부합하지 않는다고 설명함

Ⅳ. 국내·외 인공지능 윤리

4.1 인공지능 윤리 기준 마련의 필요성

인공지능은 결정과 행위를 자동화하는 기술이며, 결정과정에서 인간을 배제하기 때문에 데이터의 사용과 알고리즘 설계에 있어 선제적 윤리대응이 필수

인공지능 윤리기준이란 인공지능의 개발, 제공, 활용하는데 있어 준수해야할 범국가 차원의 기본원칙

우리 생활에서 다각도로 이용될 인공지능이 궁극적으로 인류를 이롭게하고 사회 안녕과 공익증진에 기여할 수 있도록 개발자, 제공자, 이용자 등 이해관계자 모두의 가치를 포섭하는 보편적인 윤리기준으로 기능할 수 있음

모든 인공지능은 단순히 인간에게 유용함을 넘어서 인간 고유의 성품을 보존하고 함양되도록 개발, 활용되어야 함

인공지능 윤리기준은 특히 주요 쟁점이었던 구속력과 강제력이 있는 법이나 지침과 구분하여 윤리기준의 형태로 기준을 제시했다는 점이 특징

인공지능 윤리기준의 중요성은 기업의 자율성을 존중하고 기술발전을 장려하며, 기술과 사회변화에 유연하게 대처할 수 있는 윤리담론을 형성하기 때문임

국가	주요 추진 내용
EU	-인공지능법 제안(2021.4.) ; 세계 최초 -인공지능백서 발간(2020) -신뢰할 수 있는 인공지능 윤리 가이드라인(2019)
미국	-인공지능 규제 가이드라인(2020) -인공지능 및 알고리즘 이용지침(2020)
프랑스	-'인간을 위한 인공지능' 구현에 필요한 권고사항 도출(2018)
영국	-설명 가능한 인공지능 가이드라인(2020) -공공분야 인공지능 활용 지침(2019)
일본	-인간 중심의 인공지능 사회 원칙(2018) -인공지능 개발 가이드라인(2017)

자료: 과학기술정보통신부, 신뢰할 수 있는 인공지능 실현 전략, 2021.5.13.

4.2 유럽의 신뢰할 만한 인공지능 윤리 가이드

(1) 목적

- 유럽의 신뢰할 만한 AI 윤리 가이드라인(Ethics Guidelines for Trustworthy AI)은 적법하고, 윤리적이며, 견고한 AI시스템의 설계, 개발, 이용을 위한 원칙, 요구사항 및 자율점검을 위한 평가 목록 개발

(2) 주요 내용

- AI 또는 AI 시스템에 대한 개념을 정의하고 신뢰할 수 있는 AI의 필수 3요소(적법성, 윤리성, 견고성), 7대 요구사항 등을 제안함

1) **신뢰할 수 있는 AI 3요소**

- 적법성, 윤리성, 견고성

2) **7대 요구 사항**

- 인간행위자와 감독
- 기술적 견고성과 안전성
- 프라이버시와 데이터 거버넌스
- 투명성
- 다양성, 차별금지, 공정성
- 사회적, 환경적 웰빙
- 책임성

4.3 EU 인공지능법(안)

인공지능은 4차 산업혁명 중심 기술로 인공지능 기술은 인류에게 새로운 기회와 변화와 함께 새로운 위험을 가져올 수 있다는 예측과 함께 인공지능으로 인한 새로운 유형의 피해와 침해가 발생하고 있어 전세계의 인공지능 윤리에 대한 관심은 계속 높아지고 있었음

EC는 유럽 시장에 출시 및 사용되는 인공지능의 안정성을 보장하고 EU 가입국의 시민들의 기본권을 보호하고 EU의 가치를 존중하기 위한 EU 인공지능법(안)을 제시하였음(21.4.21)

이 법은 AI가 민주적 가치와 인권, 법을 존중하면서 법률·윤리·기술적으로 견고한(robust) ,높은 수준의 신뢰할 수 있는 AI 패러다임을 성문화한 것이라는 평가

4.3.1 금지 인공지능 시스템(Unacceptable Risk AI systems)

- **EU의 가치, 인간의 존엄성, 자유, 평등, 민주주의, 기본권을 침해하고 사람들의 안전, 생계 및 권리에 대한 명백한 위협으로 간주되는 인공지능 시스템은 금지됨**
- Unacceptable Risk AI systems 4가지 범주로 구분하고 있는데, (1) 인간의 행동조작, (2) 취약성 표적, (3) 공공기관의 사회적 신용평가, (4) 실시간 원격 생체 신원 확인이 있음
- 특히, 형사사법시스템에서 고려할 부분이 (4) 실시간 원격 생체 신원확인 부분임

- 경찰 등이 공개된 장소에서 실시간으로 생체정보를 활용하여 신원확인을 하는 인공지능 시스템을 금지하고 있으며, 예외적으로 실종아동 수색, 테러 대응 등 특정범죄인 검거를 위해 사법당국의 사전승인을 받은 경우는 제외됨

4.3.2고위험 인공지능 시스템(High-Risk AI-Systems)

- 고위험 인공지능 시스템은 (1) 제품자체 또는 제품의 안전 요소, **(2) 기본권에 영향을 미칠 수 있는 인공지능 시스템**을 구분하고 있음
- 가. 생체 인식·분류: 실시간 또는 사후적으로 사람의 생체정보를 활용하여 신원확인을 수행하는 인공지능시스템
- 나. 법집행: 경찰 등이 법 집행 목적으로 인공지능시스템을 이용하는 경우
- 범죄, 재범 및 잠재적 피해자 될 위험을 평가하기 위해 사용되는 인공지능시스템
- 거짓말 탐지기 및 유사한 도구로 사용하거나 사람의 감정 상태를 감지하기 위해 사용되는 인공지능 시스템
- 지역순찰, 감시 등에 자원을 할당하기 위하여 범죄발생 가능성 및 사회적 불안을 예측하기 위해 사용하는 인공지능시스템
- 범죄수사 ·기소 단계에서 증거의 신뢰성을 평가하기 위해 사용하는 인공지능시스템

4.3.2 고위험 인공지능 시스템(High-Risk AI-Systems)

- 고위험 인공지능 시스템은 (1) 제품자체 또는 제품의 안전 요소, (2) 기본권에 영향을 미칠 수 있는 인공지능 시스템을 구분하고 있음

- 법집행: 경찰 등이 법 집행 목적으로 인공지능시스템을 이용하는 경우
- 프로파일링 기반으로 실제 또는 잠재적 범죄행위의 발생 또는 재발을 예측하거나, 자연인 또는 단체의 특성(특징) 또는 과거의 범죄행위를 평가하기 위해 사용하는 인공지능시스템
- 범죄행위의 조사 ·탐지 ·기소 과정에서 개인의 프로파일링을 위해 사용하는 인공지능시스템
- 데이터에서 알려지지 않은 패턴을 식별하거나 숨겨진 관계를 발견하기 위해 다양한 데이터 소스 또는 대규모 데이터셋을 조사하여 범죄분석에 사용하는 인공지능시스템

- 사법업무: 사실 및 법률 조사, 법률해석 및 적용 지원 등 법정에서 판사를 지원하는데 사용되는 AI 시스템

4.3.3 고위험 인공지능시스템 공급자와 사용자 의무

1. 공급자 의무

(1) 위험관리 시스템
(2) 고품질 데이터셋
(3) 품질관리 시스템
(4) 투명성 보장
(5) 인간의 관리 및 감독
(6) 견고성, 정확성 및 사이버 보안
(7) 적합성 평가
(8) 모니터링 및 보고

2. 사용자 의무

(1) 사용자 지침에 따른 적절한 기술적, 관리적 조치
(2) 시스템 운영 모니터링
(3) 심각한 사고, 오작동 또는 국가적 차원 위험 모니터링, 공급자 등에게 통지 및 시스템 중지
(4) 자동 생성된 로그 관리 보관

4.4 미국의 인공지능 윤리

- 신뢰성 있는 인공지능 기술개발 중심 국가 전략정책 아래 주요 기업을 중심으로 윤리적 인공지능 실현을 위한 자율규제를 마련하도록 하고 있음
- 미국은 2020년 인공지능 신뢰확보 10대 원칙(투명성, 공정성 등)을 내용으로 하는 규제가이드라인을 발표

- **투명성(Transparency)**

- 법은 전통적으로 정부의 업무와 특히 사법 제도의 행정에 있어 투명성을 매우 중요하므로 개발 프로세스와 방법들이 불분명하다는 지적이 있음

- **편향성(Bias)**

1 미국에서는 역사적으로 특정 인종집단이 범죄 집단에 과도하게 대표되는 경향이 있음. 이런 현실로 인해 따라서 과거 데이터로 훈련된 알고리즘 시스템은 편향적 결과를 도출할 가능성이 높음

1 기존 평가 도구는 불평등한 결과에 다양한 방식으로 접근하지만, 평가의 방법에 대한 효과에 대해서는 거의 알려진 바가 없음을 지적함

- **설명 가능성(Interpretability)**

- 대부분의 정부 결정(특히, 형사 집행)은 법적 분석과 정당성을 수반해야 함. 하지만 알고리즘은 일반인이 이해할 수 있는 설명을 제공하지 못할 수도 있음

- 기계 학습의 특성은 오늘의 케이스와 내일 유사한 케이스에 서로 다른 결과를 도출할 수 있으며, 이는 인공지능의 결과를 해석할 수 있으며 어떤 모순이나 오류도 설명될 수 있어야 한다는 것을 확실히 하는 것이 중요함

- 입법자들이 알고리즘 설명의 장점, 필요성, 한계를 이해할 수 있고, 기술적으로 실현가능한 정책의 초석이 될 수 있는 프레임워크 작업을 위한 학제 간 (법조인, 법학자, 컴퓨터 분야의 전문가, 과학자 등)의 연구가 필요함

Ethics and Governance of Artificial Intelligence Initiative: Algorithms and Justice
Ethics and Governance of AI | Berkman Klein Center (harvard.edu)

4.4.1 미국의 인공지능 윤리의 법적 정의

- **S.1705 - AICT Act of 2021**
  SEC. 3. DEFINITIONS.

(2) ARTIFICIAL INTELLIGENCE ETHICS.—The term “artificial intelligence ethics” includes the quantitative analysis of artificial intelligence systems to address matters relating to the effects of such systems on individuals and society, such as matters of fairness or the potential for discrimination

- 인공 지능 윤리는 "인공 지능 윤리"라는 용어에는 공정성 문제 또는 차별 가능성과 같은 개인 및 사회에 대한 이러한 시스템의 영향과 관련된 문제를 해결하기 위한 인공 지능 시스템의 정량적 분석이 포함됨

- 이 법의 인공지능 윤리 정의는 표면적으로 보면 매우 좁게 정의되어 있다는 평가를 받고 있는데, 이는 미국의 AI 국가 전략과 관계가 있는 것으로 판단됨

4.5 국내 인공지능(AI) 윤리

과학기술정보통신부와 정보통신정책연구원은 대통령 직속 4차산업혁명위원회 전체회의에서 인공지능시대의 바람직한 인공지능 개발·활용 방향을 제시하기 위한 사람이 중심이 되는 인공지능(AI) 윤리기준」(2020.12.23.)을 마련

「‘인간성을 위한 인공지능(AI for Humanity)’을 위해 인공지능 개발에서 활용에 이르는 전 과정에서 고려되어야 할 기준으로 3대 기본원칙을 제시함. 그리고 3대 기본원칙을 실천하고 이행할 수 있도록 인공지능 전체 생명 주기에 걸쳐 충족되어야 하는 10가지 핵심 요건도 제시함

구 분	키워드	비 고
3대 기본원칙	- 인간 존엄성 원칙 - 사회의 공공선 원칙 - 기술의 합목적성 원칙	- 인공지능 개발 및 활용 과정에서 고려될 원칙
10가지 핵심 요건	- 인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성	- 기본원칙을 실현할 수 있는 세부 요건

자료: 관계부처 합동, 인공지능(AI) 윤리기준, 2020.12.23.

Ⅴ.형사정책적 시사점

- 인공지능 윤리 가이드에 따른 인공지능 개발 및 활용
- 인권 및 기본권 침해 문제
- 법제도 정비
- 윤리 교육

1. 인공지능 윤리 가이드에 따른 인공지능 개발 및 활용

(1) 신뢰할 수 있는 AI 개발 및 이용

- 적법성, 윤리성, 견고성 있는 AI 개발 및 이용

(2) 기술적 견고성과 안정성

- AI 개발 시, 위험예방을 위한 접근방식이 존재하고 의도한 대로 신뢰성 있게 작동하는 한편 의도치 않고 예상하지 못한 위험을 최소화하고, AI 이중 용도와 같은 의도치 않은 사용과 악의적 행위자에 의한 시스템 남용 가능성을 고려해 이를 완화, 방지하기 위한 조치를 취해야 함

(3) 프라이버시

- 개인정보 존중, 데이터의 질과 무결성, 데이터 접근성

(4) 투명성

- 추적 가능성, 설명가능성, 의사소통

(5) 편향성, 공정성

- 불공정한 편향 방지, 감시프로세스 시행, 다양한 분야의 전문가 인력 의견 확보

(6) 책임성

- 기본권을 포함한 윤리 원칙에 맞지 않은 AI 의 문제에 대해 명시적으로 그 문제를 인정하고 평가할 수 있어야 하며, 윤리적으로 용인할 수 없는 경우, 기본권이 침해되는 경우에는 AI 개발, 이용, 배포 등을 금지해야 함
- 부당한 피해에 대해서는 피해자 구제를 가능케 하도록 채널과 프로세스를 갖추고 있어야 함

(7) 인간행위와 감독

- AI 시스템은 인간의 자율성과 의사결정을 뒷받침하는 역할로 사용자의 주체성을 인정하며, 인간의 인공지능 감독을 허용할 수 있어야 함
- 인간이 AI 시스템 감독권이 적다면 보다 엄격한 시험과 규제 거버넌스가 필요함
- 미국 위스콘신주 대법원(Wisconsin vs. Loomis)은 2016년 법원이 피고인들의 재범 위험성을 평가하면서 인공지능 알고리즘(컴파스)에 전적으로 의존했다면 위법이지만, 보조 수단으로 활용했을 때는 적법하다고 판결하였음

대법원은 제한사항과 주의사항을 준수할 경우 순회법원이 선고 시 COMPAS 위험평가를 고려한 것은 피고의 적법한 절차권리를 침해하지 않는다고 결론 내림

1.1 인공지능 윤리 가이드에 따른 인공지능 개발 및 활용
- 고위험 인공지능 운용・이용에 따른 의무 ・ 권리

- 국내 신뢰할 수 있는 인공지능 실현 전략(안)의 고위험 인공지능 서비스의 운용 및 이용단계에 따른 주체별 의무 및 권리(안)에 따르면, 인공지능을 활용한 업무처리 여부에 대해 이용 거부를 할 수 있도록 서비스 제공 전 해당 인공지능의 활용 여부 고지를 의무화하겠다고 함

- 하지만 현재 국내 형사사법에서의 인공지능 활용에 있어 대부분이 EU의 인공지능법(안)의 고위험 인공지능 시스템에 해당하며, 이런 인공지능 시스템 서비스가 운용되고 있는지에 대해 국민들이 알고 있는 경우가 많지 않은 것이 현실이며, 이런 현실에 비추어 보아 사전고지, 이용거부, 이용에 대한 설명 요청, 결과 설명, 이의제기에 해당되는 프로세스를 이용할 가능성이 현저히 낮아 보임

- 이런 문제는 이미 국내 시민단체 등에서 문제가 제기된 바 있음
- 실현가능한 고위험 인공지능 서비스의 운용/이용단계에 따른 주체 별 의무/권리(안)이 필요함

< 고위험 인공지능 서비스의 운용/이용단계에 따른 주체별 의무/권리(안) >

※ '자동화된 의사결정'에 대한 동 의무/권리는 △「유럽 개인정보보호법」(GDPR, '18.5~)은 기 제도화, △우리나라는 「개인정보보호법」('21.1.6일 입법예고)에서 제도화 추진 중

사람이 중심이 되는 인공지능을 위한 신뢰할 수 있는 인공지능 실현 전략(안)

2. 인권 및 기본권 침해 문제

- 범죄 예방, 효과적 용의자 검거 등을 목적으로 하는 공적 영역에서 인공지능 활용이 증가하면서 대중적 공감을 얻고 있지만 공적영역에서의 기대 효과 중심으로만 인공지능의 활용을 인정하게 되면 이 후 발생할 심각한 기본권 침해, 인권 침해의 문제가 발생할 수 있음
- 인공지능 윤리기준 거버넌스의 공통적 원칙, EU의 인공지능법(안)에서 금지, 고위험 인공지능으로 경찰 등 형집행기관에서 사용하는 인공지능이 다수 포함하고 있는 점 등을 진지하게 고민할 필요가 있음
- 특히, 이미 외국에서는 안면인식 기술 등의 사용을 금지하는 곳도 있으며, 안면인식 기술을 이용한 인공지능 활용에 있어 명확한 기준을 법률로써 두고자 하는 곳도 있는 만큼 우리도 진지한 검토가 필요할 때임

3. 법제도 정비

- 국내 법체계와 해외 입법 동향 등을 분석하여 글로벌 스탠다드와 조화를 이루면서 우리 실정에 맞는 법제 마련이 시급함
- 최근 미국의 형사사법 인공지능 기술 중 안면인식기술에 관한 법안 등을 통해 인공지능의 형사정책적 시사점을 도출해 볼 수 있음
- 미국 상원의원 Gsv}$Fssoiv$erh$Nijj$Q ivopi}은 6464년에 위원회가 안면 인식 기술 사용에 대한 적절한 지침과 제한을 권고할 때까지 얼굴 인식 기술의 정부 사용을 유예하는 법안을 제안했으며, 이 법안은 이 문제를 연구하여 의회에 법안을 제안할 의회 위원회를 만들 것을 그 내용으로 하고 있음
- 이 법안에 따르면, 위원회의 역할은 기술이 윤리기준에 따라 개발되고, 개인의 프라이버시 권리를 침해하거나 편향된 결과를 초래하지 않도록 보장하기 위해 " 가이드라인과 규제" 을 만드는 것을 내용으로 하고 있음.
- 미 연방법 the Facial Recognition and Biometric Technology Moratorium Act of 2021 은 연방 정부의 FRT 사용을 금지하고 이 기술을 사용하는 주 및 지방 법 집행 기관에 대한 연방 지원을 제한하고, 개인의 프라이버시권에 대한 소송권을 인정함

- 최근, 미국 주 의회는 형사사법에서의 안면인식기술과 관련하여 다양한 법률을 제정하거나 법안을 발의했는데, 대부분 민간 부문보다는 정부 기관을 규제하는 것을 목표로 하고 있음
- 미국 주법상 형사사법에서의 인공지능 안면인식기술에 관한 법 크게 3가지로 나눌 수 있음

(1) 사전 법적 승인(Virginia and Pittsburgh, Pennsylvania)

- 매사추세츠와 유타 주는 a facial recognition search 전에 데이터베이스를 관리하는 주 당국에 서면 요청서를 제출할 것을 요구

(2) 영장이나 법원 명령요구

- 매사추세츠와 워싱턴에서는 사법당국이 FRT를 사용하기 전에 영장이나 법원의 명령을 받도록 함으로써 사법감독이 부과됨
- 메인 주의 경찰관들은 FRT 요청을 하기 전에 probable cause 을 충족해야 하며, 수색이나 체포를 이유로 안면 인식 일치여부를 확인하는 것을 금지함

(3) 사용 금지(오리건 주, 뉴햄프셔주, 캘리포니아주)

-뉴저지, 뉴욕, 사우스 캘리포니아도 비슷한 법을 발의하였음

4. 윤리 교육

- 인공지능윤리 교육의 필요성에 대해서는 많은 이들이 공감하고 있음
- 인공지능(또는 알고리즘) 책임과 관련하여 인공지능 알고리즘 설계자, 기술자 등의 인공지능 윤리 교육은 필요함
- 인공지능 설계를 위해 설계를 요청한 자, 인공지능을 사용할 자에 대한 인공지능 윤리 교육도 필요함
- 즉, 인공지능 설계부터 안전한 사용에 이르기 까지 윤리 교육, 윤리기준, 관련 법률에 대한 교육이 이루어질 필요가 있음

< 토 론 문 >

인공지능 윤리가이드와 형사정책적 시사점 - 토론문 -

전북대학교 법학전문대학원 조기영 교수

Ⅰ.

우리는 오늘날 알고리즘이 조종하는 자율주행 자동차를 운행하고, 번역기의 도움으로 외국어로 이메일을 쉽게 쓸 수 있게 되었습니다. 인공지능에 의해 가능해진 새로운 일상은 형법학에 대하여도 새로운 질문을 제기할 수밖에 없습니다. 인공지능은 우리 공동체뿐만 아니라 형사사법에도 변화를 가져올 것이며, 형사사법의 관점에서도 이제 인공지능을 무시하거나 소홀히 할 수 없는 상황이 되었습니다. 인공지능이 효과적·효율적인 법익보호를 가능하게 하고(**안전형법**), 자유적이고, 법치국가적이며 민주적인 객관적이고 일관성 있는 형법의 적용을 보장할 수 있는지(**자유형법**) 질문을 해야 할 때입니다. 이러한 상황에서 김희정 교수님께서 발표해 주신 '인공지능 윤리가이드와 형사정책적 시사점'은 형법학자들로 하여금 진지한 고민을 촉구하고 있다고 생각됩니다.

Ⅱ. 인공지능의 형사사법에의 적용영역

1. 범죄(인) 예측

경찰의 범인, 범인의 특성, 범죄의 시간, 장소 등에 대한 데이터를 기초로 하는 범죄예측 알고리즘은 범죄사회학적 전제를 가지고 있습니다(Near-Repeat-Ansatz).[1] 알고리즘을 이용한 범죄예측으로는 Hot-Spot-Idenfikation과 Crime Linking이 논의되고 있는 것 같습니다. 김희정 교수님께서 소개해 주신 한국 경찰의 '프리카스'도 여기에 해당하는 것으로 보입니다. Hot-Spot-Idenfikation에는 미국의 주들이 적용하고 있는 RTM(Risk Terrain Modeling) 시스템, UCLA와 LAPD 및 영국의 연구프로젝트인 PredPol(Predictkve Policing), 이탈리아에서 개발한 XLAW, 독일과 스위스 일부에서 이용하고 있는 Precobs(Pre Crime Observation Sytstem) 등이 알려져 있습니다.[2]

발생한 범죄와 특정 개인을 연결시켜 언제 어디에서 그 개인이 범죄를 범할 것인지 예측하는 Crime Linking은 연속범(Serientäter)의 반복적 범행 특성을 알고리즘을 통해 분석한다고 합니다. 영국 캠브리지 대학이 개발한 HART(Harm Assessment Risk Tool), 이탈리아 경찰에서 강도범행 예측을 위해 사용하는 KeyCrime 소프트웨어가 여기에 속한다고 합니다.

2. 수사절차

1) Lukas Staffler, Künstlliche Intelligenz und Strafrechtspflege – eine Orientierung, ZIS 4/2020, S. 167.
2) Lukas Staffler, a. a. O., S. 167-168.

위에서 언급한 바와 같이 시민의 안전을 위한 경찰작용은 인공지능의 도움을 많이 받을 수 있을 것으로 생각됩니다. 경찰은 인공지능을 통해 더욱 많은 정보를 수집하고 위험을 예측하고 예방할 수 있을 것입니다. 반면, 수사절차에서는 범죄의 혐의가 인정되면 수사기관은 수사를 개시·진행해야 한다는 '수사강제주의'가 적용됩니다.[3] 인공지능 시대에 수사기관이 방대한 정보를 수집할 수 있게 됨에 따라 오늘날 수사기관의 과제는 다양한 디지털 정보의 수집이 아닌 이들 정보의 선별이라는 과제에 직면해 있다고 보입니다.

2019년 독일 Nordrhein-Westfalen 법무부는 인공지능을 위하여 아동포로노그라피에 투쟁하기 위해 인터넷상의 광대한 자료 중 범죄와 관련된 정보를 수집하는 소프트웨어를 사용하고 있다고 합니다.

3. 공판절차

(1) 증거조사

공판절차에 알고리즘을 적용하는 것, 즉 증거조사 내지 재판에 알고리즘을 사용할 수 있을지는 어려운 문제에 속합니다. 먼저, **증거조사**와 관련하여 공판절차에는 직접주의 원칙이 적용된다는 점을 고려하여야 합니다. 수사기관이 수집한 디지털화된 증거자료도 직접주의원칙에 따라 법관이 이를 사본의 형태로 읽을 수 있어야 합니다. 읽을 수 없는 디지털증거가 증거로 제출된 경우 이를 어떻게 처리해야 할지 문제가 됩니다. 법관이 스스로 확인할 수 없는 알고리즘에 의해 결정된 결론을 수용하는 것이 직접주의원칙에 비추어 어느 범위에서 가능할 수 있는지가 문제가 될 것입니다.

김희정 교수님께서 소개해 주신 법원이 피고인들의 재범위험성을 평가하면서 인공지능알고리즘(컴파스)에 전적으로 의존했다면 위법이지만, 보조수단으로 활용했을 때는 적법하다고 판시한 2016년 미국 위스콘신주 대법원의 판결을 - 물론 이 판결은 재범위험성의 측정에 관한 것이기는 하지만 - 하나의 기준으로 고려하여 법관이 확인할 수 없는 알고리즘의 결론을 법관이 심증형성의 주된 기초로 삼을 수는 없지만 보조자료로는 활용할 수 있다는 방안을 검토해 볼 수 있을 것 같습니다.

(2) 법원의 재판

몽테스키외나 베버는 법관이 포섭기계여야 한다고 생각하였지만, 오늘날에는 형벌구성요건에 의한 책임의 존부와 범위 판단을 수학적으로 정하는 것은 불가능하고 법관의 재량이 작용한다는 점에 의문이 제기되지 않고 있습니다. 법관은 피고인, 증인 등의 진술을 음미하고 그 신빙성을 판단하며 경우에 따라 당해 사건의 정치적, 사회적, 철학적, 경제적 상황과의 관련성을 고려하는 것으로 보입니다. 우리는 종종 법관의 판단에 매우 불만족하기도 합니다.

우리는 머지않아 법원에 근무하는 법관들을 인공지능에 기반을 둔 시스템으로 완전 대체하자는 주장을 목격하게 될지도 모릅니다. '헌법과 법률이 정한 법관'에 의한 재판이라는 헌법적 요청을 무시할 수는 없더라도, 인공지능이 전혀 의문의 여지없는 사실관계를 완벽하게 확정하는 것이 더 이상 순수한 가설이 아니라면 이러한 주장은 더욱 힘을 얻을 수 있을 것입니다.

인공지능이 알고리즘에 따라 판결을 선고하는 광경을 목격하기는 쉽지 않겠지만, 인공지능이 피고인의 범행수법 및 범행의 결과, 피고인의 과거 전과 및 생활환경, 피고인에 대한 재판기간, 구금에 따른 자유박

---

3) 신양균/조기영, 형사소송법, 박영사, 2020, 90쪽. 다만, 범죄의 혐의는 인정되지만 수사의 필요성의 내재적 한계로서, 실질적으로 입건의 필요성이 없거나 입건하는 것이 국익이나 공익에 반하는 경우 또는 피해가 보호가치가 없거나 극히 희박한 경우 등 수사의 필요성이 없는 사건에 대하여는 수사강제주의에 대한 예외를 인정할 수 있을 것이다. 2020년 개정형사소송법 제196조는 종래 범죄의 혐의가 있다고 사료하는 때에는 범인, 범죄사실과 증거를 '수사하여야 한다'에서 '수사한다'로 그 문언을 변경하였다.

탈 기간 등을 고려하여 적정한 양형의 기준을 제시해 주고 이에 따라 형사절차를 신속하고 효율적으로 종결하는 시대가 올지도 모르겠습니다.

4. 위험예측

인공지능은 구속전 피의자심문 및 구속적부심사에서 피의자에 대한 구속 또는 석방 여부, 피고인에 대한 보석허가 여부, 피고인에 대한 선고유예·집행유예의 선고 여부, 보안처분의 부과 및 수형자에 대한 가석방 여부 판단을 할 때 피의자 또는 피고인 및 수형자의 장래의 위험성 예측에 효과적으로 활용될 수 있을 것으로 생각됩니다. 독일과 스위스 취리히에서 형집행 및 보안처분 실시에 사용하고 있는 FORTES(Foresisches Operationalisierte Theraphie-Risiko-Evaluationssystem), 미국에서 사용하고 있는 PSA(Public Satety Assesment)와 COMPASS(Correctional Offender Management Profiling for Alternative Sanctions) 등이 여기에 해당합니다.

위험예측은 물론 형사사법절차 전반에서 인공지능이 활용되기 위해서는 김정희 교수님께서 지적해주신 것처럼 알고리즘의 편파성, 투명성 결여, 오용 등의 문제점을 사전적으로 통제할 수 있는 방안이 확보되는 것을 전제로 하여야 할 것입니다. 이러한 의미에서 국내외의 인공지능 윤리기준을 소개해 주시고 그 시사점을 제공해주시는 김희정 교수님의 발표는 현재의 상황에서 형사사법절차에 대한 연구를 행하는 형법학자들에게 소중한 가르침이라고 생각됩니다.

Ⅲ.

인공지능은 형사사법의 패러다임의 전환을 가져올지도 모릅니다. 기술의 발전은 고전적 원칙을 고수하는 형법학에 비판적 시각을 제공해 줄 수 있습니다. 그러나 이는 동시에 인간의 기본권에 대한 중대한 위협이 될 수 있음은 물론입니다.

새롭고 발전하는 기술 앞에서 법학자가 말할 수 있는 것은 법적 원칙입니다. 2018년 유럽에서 공표된 '사법시스템에 인공지능을 활용하기 위한 윤리장전'이 주목을 받는 이유도 여기에 있다고 생각됩니다. 윤리장전의 배후에 있는 다섯 가지 원칙은 다음과 같습니다.[4)]

- 인권존중의 원칙
- 비차별의 원칙
- 성능(Qualität)과 안전의 원칙
- 투명성, 독립성 및 공정의 원칙
- 인간의 개입 보장의 원칙

이러한 원칙을 바탕으로 형사사법에의 인공지능 활용 방안을 검토해 나가는 것이 이후 형법학자들의 임무가 될 것 같습니다. 공동체가 알고리즘으로 전환하는 것이 형법의 종말을 의미하는 것은 아닌지 고려해 보아야 합니다. 인공지능이 자유주의적 형법을 쇠약하게 하고 결국 사망에 이르게 하거나 복지국가적 안전형법의 종말을 의미하는 것이 아니라 기술에 대해 개방적이고 인간의 기본권에 적합한 미래의 형법의 반석이 될 수 있어야 할 것입니다. Joachim Hruschka가 강조하고 있듯이 이성적 논증을 활용하고 마법과 같은 악마의 주문이 되는 것을 피하기 위해서는 형법학은 항상 지적으로 성실해야 할 것입니다.[5)]

4) Lukas Staffler, a. a. O., S. 173.
5) Hruschka, Strafrecht nach logisch-analytischer Methode, Berlin/New York: De Gruyter, 1988 , S. XVIII.

Ⅳ.

마지막으로 형법학자들에게 익숙치 않은 인공지능의 형사정책적 문제에 관한 발표를 해주셔서 많은 배움의 기회를 주신 김정희 교수님께 감사의 말씀을 드리며, 토론자의 역할 중의 하나인 질문을 드리는 것으로 이번 토론을 마치고자 합니다.

교수님께서 소개해 주신 우리 정부가 추진하고 있는 인공지능의 형사사법에의 활용방안은 주로 '안전형법'을 염두에 둔 것으로 보입니다. 범죄 발생 확률을 예측해주는 '프리카스', 법무부의 인공지능을 이용한 전자감독시스템, 인공지능 보호관찰 등이 바로 그러합니다. 범죄를 예방하고 재범을 방지하는 것은 공동체의 주요한 과제 중의 하나입니다. 그러나 다른 한편, 수용자의 재사회화를 통하여 석방 이후의 수용자의 자유로운 생활을 가능케 하면서 사회를 보호하기 위한 노력이 소홀히 될 수 없음은 물론이다. 독일 형법이론에서 응보형이론이 강조되고 특별예방에 대한 관심이 적어 범인의 재사회화를 위한 기본법인 형집행법이 1977에 와서야 제정된 것[6]은 이점을 잘 보여 줍니다. 최근 핀란드에서는 교도소에 디지털 시스템을 도입하여 수용자의 사회복귀에 도움을 준다고 합니다. 컴퓨터를 이용하여 가족이나 친지들과 소통할 수 있으며, 사회복귀에 도움이 되는 인터넷 학습을 할 수 있다고도 합니다.

현재 정부의 형사사법에서의 인공지능 활용 방안이 범죄예측·예방에 초점을 두고 있는 것으로 느껴지는 상황에서 향후 특별예방의 기능의 효율화를 위하여 인공지능을 활용할 수 있는 방안에는 무엇이 있을 수 있을지 여쭙고자 합니다.

감사합니다.

---

6) Roxin/Greco, Strafrecht AT I, 5. Aufl. 2020, § 3, Rn. 9.

세션 II

< 발 표 문 >

AI를 활용한 구속기준의 재정립

동의과학대학교 경찰행정계열 교수 김병수

Ⅰ. 문제의 제기

사회의 많은 분야에서 머신러닝 및 딥러닝 기술을 바탕으로 한 인공지능(AI)이 빅데이터 속에서 숨겨진 패턴이나 상관관계를 찾아내고 있다. 이를 바탕으로 지금까지 알지 못했던 새로운 지식을 발굴하여 앞으로 어떠한 일이 발생할 것인지를 전망하고 있다. 이러한 인공지능의 예측은 통계적으로도 유의미한 결과를 내놓고 있어 금융, 마케팅, 운송, 의료, 공공서비스 등 사회전반에 활용되고 있다. 이러한 AI의 높은 예측력은 범죄예방분야에서도 활용되고 있는데, 그것이 바로 범죄예측 시스템이다. 이미 범죄예측 시스템을 도입한 선진국에서는 범죄율이 감소하는 등 가시적인 성과를 나타나고 있다.[1)]

그러나 범죄예측 시스템은 경찰학 등과 같은 경험과학에서는 활발히 연구가 되고 있으나 규범학인 형사법분야에서는 상대적으로 연구가 미진한 것이 사실이다. 특히 범죄율나 재범률과 같이 가시적인 성과가 나오는 분야에서는 범죄예측 시스템을 이용한 연구가 활발히 이뤄지고 있으나, 양형기준이나 구속기준과 같은 규범적 판단기준을 정립하는 데에는 범죄예측 시스템을 이용한 연구는 제대로 이뤄지지 않고 있다.

구속기준이나 양형기준은 '유전무죄 무전유죄', '고무줄 양형', '전관예우', '유전불구속 무전구속'로 대표되는 국민의 사법에 대한 불신을 해소하기 위하여 시급히 해결하여야 할 과제이다. 양형위원회가 설치되고 객관적이고 투명한 양형판단을 위해서 비록 권고적 기준에 불과하지만 양형기준이 제정되고 있다.[2)] 그렇다면 양형기준 못지않게 국민으로 하여금 사법의 정당성에 대한 의구심을 갖게 하는 구속기준도 재정립하는 것이 시대의 요청인 것이다. 하지만 양형기준에 비해 구속기준에 대한 사회적 지원은 전혀 이뤄지지 않고 있다. 그럼에도 불구하고 구속영장의 발부와 기각을 둘러싼 논란은 끊이질 않고 있다.[3)] 최근 검찰의 구속영장 청구를 기각한 판사를 적폐로

---

1) Zach Friend, Predictive Policing: Using Technology to Reduce Crime, FBI Law Enforcement Bulletin(Feb. 8, 2018).; C. Francescani, "NYPD Expands Surveillance net to Fight Crime as Well as Terrorism," Reuters, June 21, 2013.; Jennifer Bachner, Predictive Policing: Predictive Policing: Preventing Crime with Data and Analytics, Washington, DC: IBM Center for The Business of Government, 2013.; Jeffrey S. Paul and Thomas M. Joiner, Integration of Centralized Intelligence with Geographic Information Systems, Geography and Public Safety, Oct. 2011, at 7.

2) 양형기준을 통하여 양형과정에 국민의 건전한 상식을 반영함으로써 국민이 신뢰할 수 있는 공정하고 객관적인 양형을 실현하는 것을 목표로 설립된 2007년 제1기 양형위원회는 지금은 제7기 양형위원회(2019. 4. 27 ~)로 활동하고 있다.
출처 : https://sc.scourt.go.kr/sc/krsc/intro/greet/purpose.jsp

3) "판사 '고무줄 영장'… 검찰서 행패는 기각, 법원 난동은 구속", 오락가락하는 법원의 구속영장 발부기준, 같은 판사가 맡은 비슷한 사건인데 구속영장 발부는 정반대, 조선일보 2019.05.29.

몰고 주소와 가족사진을 인터넷에 공개하는 '신상털이'[4]와 같은 여론에 의한 테러가 일상화되고 있다. 영장청구와 기각을 둘러싼 검찰과 법원의 갈등이 이제는 여론의 분열과 함께 사회의 갈등으로 확산되고 있다.

이것은 '증거인멸', '도주우려' 등 현행 구속사유(형사소송법 제70조)가 너무 추상적이어서 구체적 사건에서 정확한 구속기준이 뭔지 알기도 어렵고 예측하기도 어렵기 때문이다. 게다가 법률전문가인 검찰과 법원도 국민들이 공감할 수 있는 명확한 판단기준을 제시하고 못하고 있다. 이 때문에 국민들은 어떤 경우에 구속이 되는지를 몰라 인신구속의 두려움을 해소하기 위하여 비싼 수임료를 부담해서라도 전관출신의 변호사를 찾고 있는 것이 현실이며, 재벌총수와 정치인 등 돈 있는 사람과 힘없고 돈 없는 서민의 구속기준이 다르다고 생각하여 '유전불구속 무전구속'이란 말까지 하여 왔다. 국민들의 법률생활을 불안정하게 하고 예측가능성을 곤란하게 하며 형평성에 의구심까지 갖게 하는 구속기준을 재정립하는 것은 시대의 요청인 것이다.

따라서 본 논문에서는 범죄예측 시스템의 현황을 살펴보고(Ⅱ), 현행 구속기준의 문제점을 검토하고(Ⅲ), AI를 활용한 구속기준의 재정립 방안을 검토하고자 한다(Ⅳ).

Ⅱ. AI를 이용한 범죄예측 시스템의 현황

범죄통계에 따르면 범죄를 신고하는 전화의 절반이 도시 전체 주소 중에서 3.5%에 해당하는 특정 지역에서 걸려 왔고,[5] 범죄가 많이 발생하는 장소는 전체 도시의 도로 중에서 4.5%에 해당하는 특정 도로에서 발생하였고,[6] 전체 범죄자 중의 10%에 해당하는 범죄자가 전체 강력범죄의 50%를 저지른다고 한다.[7] 또한 연쇄살인범의 상당수는 어린시절 동물학대와 방화와 같은 일탈행위를 한 경험이 있으며,[8] 부모나 주위 사람들로부터 학대를 당한 경험이 있고[9] 그에게 학대를 행한 사람은 대부분 그의 부모였다.[10] 이처럼 범죄자와 범죄의 시간과 장소 등에는 일정한 경향이 발견되면서 범죄예측의 연구가 시작되었다. 오늘날에는 인공지능이 빅데이터에서 일정한 패턴

출처 : http://news.chosun.com/site/data/html_dir/2019/05/29/2019052900217.html

4) "두달 만에 "영웅"서 "부역자"로… 판사를 여론재판", 조선일보 2017.06.05.
출처 : http://news.chosun.com/site/data/html_dir/2017/06/05/2017060500066.html

5) Sherman, L.W., Gartin, P.R., & Buerger, M.E., "Hot Spots of Predatory Crime: Routine Activities and the Criminology of Place," Criminology, 27, 1989, pp.27-55.

6) 시애틀에서 14년간의 범죄 데이터를 분석한 Harthorne은 범죄의 절반이 도시 전체 도로의 4.5%에서 발생하는 것을 밝혀냈다. Harthorne M., City : A Few streets Responsible for Half of seatle's Crime; komonews.com/archive/city-a-few-streets-responsible-for-half-of-seatles-crime.

7) Nath, S.V., Crime Pattern Detection Using Data Mining, Oracle Corporation, 2006.

8) Wright, J., & Hensley, C. (2003). et. al.

9) Holmes, S. T., Tewksbury, R., & Holmes, R. M. (1999). Fractured Identity Syndrome. Journal of Contemporary Criminal Justice, 13(3), p.265 ; Whitman, T. A., & Akutagawa, D. (2004). Riddles in serial murder: A synthesis. Aggression and Violent Behavior, 9, p. 699.

10) Singer, S. D., & Hensley, C. (2004). Applying Social Learning Theory to Childhood and Adolescent Firesetting: Can It Lead to Serial Murder?. International journal of offender therapy and comparative criminology: official organ of the Association for Psychiatric Treatment of Offenders (APTO), 48(4), p.465.; Wright, J., & Hensley, C. (2003). From Animal Cruelty to Serial Murder: Applying the Graduation Hypothesis. International Journal of Offender Therapy and Comparative Criminology: official organ of the Association for Psychiatric Treatment of Offenders(APTO), 47(1), p.88.

을 찾아내고 이를 근거로 군집·최적화·예측·분류 등을 위한 판단규칙을 스스로 학습하는 머신러닝을 사용하여 범죄를 예측하고 있다.[11)]

범죄예측 시스템의 방법은 범죄 발생의 위험성이 높은 장소나 시간을 예측하는 '범죄 예측', 범죄 피해의 가능성이 높은 개인이나 집단을 전망하는 '범죄 피해자 예측', 범죄를 저지를 위험성이 높은 개인을 식별하는 방식 '범죄자 예측' 등 크게 3가지로 구분한다. 한편으로는 범죄예측 시스템을 범죄자 및 피해자의 정보를 수집하여 예측하는 인적 기반 범죄예측시스템과 장소와 같은 사회환경에 관한 정보를 활용하는 장소 기반 범죄예측시스템으로 나누기도 한다.[12)] 이 중에서 과거의 수사기록, 범죄기록, 재판기록, 판결문 등을 머신러닝을 이용한 인공지능으로 분석하여 재범의 위험성을 측정하는 연구[13)]는 인적 기반 범죄예측시스템에 속한다. 아래에서는 각국의 범죄예측 시스템의 현황을 살펴보기로 한다.

1. 미국의 PredPol과 Crime Prediction and Prevention

2008년 미국 LA 경찰청(LAPD)에서 예방적 경찰활동(Predictive Policing)으로 특정 범죄에 대한 예측시스템인 프레드폴(PredPol)을 개발하였다. 프레드폴은 머신러닝 알고리즘을 이용하여 범죄 유형, 범죄 발생 위치와 시간에 대해 예측하여 언제 어디서 범죄가 발생할지를 예측한다. 이러한 머신러닝 알고리즘은 과거의 범죄에 대한 자료 즉 데이터를 분석하면 언제 어디서 어떠한 범죄가 발생할 가능성이 가장 높은지를 예측할 수 있다. 이를 바탕으로 범죄 발생 가능성이 높은 지역에 경찰력을 우선 투입하는 방법으로 경찰의 순찰경로를 추천하는 등의 기능을 제공한다.[14)] 프레드폴은 다양한 범죄 영향 효과를 표현하기 위해 범죄유발 요인들의 효과를 시공간적으로 반영하는 시공간 분석(Spatial-temporal Analysis) 통계기법이 주로 사용된다. 이 기법 중 기술적으로 가장 많이 활용되는 대표적인 분석 모델은 인근 반복 모델(Near Repeat Model)과 위험영역모델(Risk Terrain Model)이다.[15)] 이들 시스템들의 특징은 지역의 환경요인에 대한 다양한 정보를 수집하고, 도시 및 지역별 광역 범죄분석 체계를 독립적으로 운영함으로써 예측의 정확성을 향상시키고 지역 범죄에 더욱 유연하게 대처할 수 있는 것이 특징이다.[16)]

IBM은 공공안전을 위한 효과적 경찰활동을 위해 '범죄예측 및 예방 솔루션'(Crime Prediction and Prevention)을 개발했다. 범죄와 잠재적으로 상관관계가 있다고 여겨지는 사건보고, 범죄유형, 목격자, 피해자, 용의자, 지도, 날씨, 온도, 월 및 주, 휴일, 축제일 또는 월급일, 소셜 미디어의 게시물 등의 다양한 데이터를 분석하여 다가올 경찰활동의 시간활용을 최적화하는 것이 그 목적이다.[17)]

---

11) 이태억, "[이태억의 과학에세이] 스몰데이터 AI와 뉴턴의 사과", IT조선, 2018. 7. 7., http: //it.chosun.com/site/data/html_dir/2018/07/06/2018070602627.html

12) 윤상현/신상화, 앞의 글, 251-253면.

13) 김태완, "데이터 마이닝 기법을 활용한 청소년 재범 관련요인 연구", 교정복지연구(제35호), 한국교정복지학회 (2014. 12); 이수정, "성인 판결전 조사 활용, 재범 위험성 평가도구 개발", 동향/연구보고서, 법무부(2010).

14) PredPol available from http://www.predpol.com/

15) 윤상현/신상화, 앞의 글, 251-253면.

16) 윤상현, "범죄예측 기술 어디까지 왔나?", 치안정책리뷰 제 60호, 치안정책연구소. 2018, 5-8면.

17) IBM Public Safety and Policing available from https://www.ibm.com/industries/government/public-safety/crime-prediction-prevention

2. 영국의 OASys와 NDAS

영국 런던경찰청은 사물인터넷(Internet of Things)을 활용한 범죄예측 시스템 OASys(Offender Assessment System)를 활용하고 있는데, 이 시스템은 주로 조직범죄 데이터베이스에 있는 우범자들의 범죄기록과 SNS 기록 등을 바탕으로 범죄가능성을 예측하고 있다.[18] 또한 국가정보분석 솔루션(NDAS: National Data Analytics Solution)은 대량의 데이터를 검토하여 범죄예측에 유효한 지표를 구축하였는데, 이를 통해 영국 경찰은 범죄 발생 가능성이 크다고 여겨진 사람들을 대상으로 상담 지원을 계획하고 있다. 이 프로젝트는 특정한 사회 관계망(예를 들어 폭력 서클, 경기장 훌리건, 조직범죄 등)에 속한 사람들과 이들이 범죄를 저지를 가능성 사이에 높은 상관관계가 있다는 전제를 바탕으로 하고 있다.[19] 영국 NDAS의 범죄예측 시스템은 결과적으로 장래에 범죄의 발생 가능성이 큰 장소와 시간을 지목하는 것보다 범죄인이 될 가능성이 높은 '특정인'에 주목한다는 점에 특징이 있다.[20]

3. 일본과 중국

범죄예방에 인공지능과 빅데이터 기술을 활용하고 있는 것은 일본과 중국에서도 마찬가지이다. 일본에서는 통화하는 사람의 음성을 통하여 심리 상태를 분석하고 이를 통해 보이스피싱을 예방하는 시스템을 개발하였다. 이 시스템은 혐의자로부터 의심스러운 단어가 사용되었거나 피해자의 집중력이 떨어진 정황이 있으면 이를 경찰 등에 알림으로써 보이스피싱 피해를 방지하고 있다.[21] 또한, 카나가와현 경찰청은 과거 10년간의 데이터 260만 건을 바탕으로 미래의 범죄 시간 및 장소를 예측하는 카나가와 범죄예측 시스템을 개발하고 있다.[22] 중국도 사물인터넷을 이용한 드론과 스마트글라스(smart glasses)를 활용하여 치안활동을 보조하고 있다.[23] 세계 CCTV 점유율 1위 기업인 중국 카메라 업체 하이크비전(Hikvision)은 방대한 시민정보를 빅데이터와 인공지능에 적용하여 2013년 카메라를 통해 딥러닝 영상처리를 가능하게 하는 애플리케이션을 개발하였다. 이를 통해 범죄자의 안면을 인식하고, 차량의 번호판을 식별하고, 범죄의 의심이 가는 행동하는 대상자를 색출하는 등의 기능을 제공하고 있다.[24]

4. 한국의 범죄예측 시스템

---

18) 박민지/서정범, "사물인터넷을 활용한 경찰활동에 관한 연구- 그 현황과 허용성에 관한 논의를 중심으로 -", 경찰학연구 제19권 제3호, 경찰대학, 2019. 9, 101면.
19) 행정안전부/한국정보화진흥원, 월간 디지털 정부 최신 해외 정책ㆍ기술 동향 제40호, 2018, 40면.
20) 윤영석, "범죄징후 예측시스템에 대한 비판적 검토-2019년 법무부의 발표내용을 중심으로-", 과학기술법연구 제26집 1호, 2020, 142-143면.
21) 양주아/양문승, "경찰의 과학치안 활동 현황 및 전망", 한국경찰연구 제16권 제4호, 2017. 12, 154면.
22) 행정안전부/한국정보화진흥원, 월간 디지털 정부 최신 해외 정책ㆍ기술 동향 제40호, 41면.
23) 정재준, "사물인터넷(IoT)의 치안대책 활용 방안", 경찰학연구 제18권 제1호, 경찰대학, 2018. 3, 202-203면.
24) 박민지/서정범, 앞의 글, 102면.

1) GeoPros와 CLUE

범죄발생분포 및 추세, 범죄유형 등의 범죄통계자료를 이용한 추계적 분석, 차원 공간 분석 및 가상활동 객체를 활용한 범죄발생 장소 및 시간예측, 범죄유형이나 범죄발생 공간정보, 인구통계 등의 정형데이터와 주민신고 등의 비정형데이터로 이루어진 빅데이터 분석기반 범죄예측 등 국내에서도 범죄예측가능성과 관련한 문헌적 연구가 있었다.[25] 이러한 연구를 바탕으로 지오프로스(GeoPros)나 CLUE 등과 같은 위험예측모델 기반의 지능형 치안시스템들이 개발되었다. 2009년 공개된 지오프로스는 연쇄범죄에 대한 지리적 프로파일링, 용의자 색출뿐만 아니라, 범죄 위험지수, 핫스팟(hot spot) 등의 정보를 제공한다. 경찰의 순찰경로 선정 등에 이러한 정보가 활용되기도 한다.[26]

CLUE(Crime Layout Understanding Engine)는 범죄의 분포를 이해하는 도구로 2016년부터 2018년 사이에 개발된 범죄분석플랫폼이다. CLUE는 형사사법정보시스템(KICS)에 저장되어 있는 수사보고서 및 사건접수정보 등 경찰에서 보유하고 있는 정형/비정형데이터 및 경찰관서 위치, 경찰인원 등의 치안정보데이터와 52종의 공공데이터를 결합하여 범죄경향분석, 수사단서제공, 유사사건추천, 범죄예측 등의 분석결과를 제공한다.[27]

2) 법무부의 범죄징후 예측시스템

범죄징후 예측시스템은 전자발찌와 같은 전자감독 대상자의 재범위험성을 분석하여 보호관찰관의 업무를 지원하는 시스템이다. 즉, 부착명령 청구전조사서, 판결문, 범죄사실, 면담정보 등 기존의 자료를 시스템으로 과학적으로 분석하여 집중적인 지도가 필요한 대상자를 보호관찰관에게 미리 알려 주어 선제적인 재범방지 활동을 할 수 있도록 지원해 주는 시스템이다. 근거법령으로는 특정범죄자에 대한 보호관찰 및 전자장치 부착 등에 관한 법률 제1조, 제15조 제1항, 보호관찰 등에 관한 법률 제33조 제1항, 제2항을 들 수 있다. 이러한 법무부의 범죄징후 예측시스템은 전체적으로 볼 때 경찰의 순찰활동이나 프레드폴 시스템처럼 특정 지역의 범죄발생을 예측하기보다는 '특정인'의 우발적 행동을 감시하여 그 특정인의 범죄를 방지하는데 초점이 있다.[28]

3) ETRI의 예측적 영상보안 원천기술

한국전자통신연구원(ETRI)은 인공지능 기술을 활용하여 과거 범죄통계 정보와 CCTV(폐쇄회로TV)영상을 자동 분석해 범죄 등 위험상황이 발생할 가능성을 보여주는 '예측적 영상보안 원천기술'을 개발하는 사업에 착수하였다. 2022년 완성을 목표로 하는 이 기술은 개별적인 행위자의 인

25) 임승옥 외, "범죄 발생환경, 행동패턴 및 심리정보 등 융합정보 적용형 엔트로피 필터링 예측분석기반의 실시간 범죄 예측· 예방 시스템 개발," 정보통신· 방송 연구 개발사업 최종보고서, 전자부품연구원, 2016. 10.
26) 윤상연, 앞의 글. 5-8면.
27) 윤상연, 앞의 글. 5-8면.; 한국전자통신연구원, "위험 상황 초기 인지를 위한 ICT 기반의 범죄 위험도 예측 및 대응 기술 개발," 한국전자통신연구원 사업계획서, 2018.
28) 윤영석, 앞의 글, 153면.

상착의와 행동을 분석함으로써 향후 일어날 수 있는 구체적인 범죄를 예측하는 방식이다. 또한 이 기술은 현재 상황을 먼저 분석하고 과거의 범죄 데이터를 비교해 범죄의 발생을 예측하는 방식으로 현재 선진국에서 진행 중인 통계적 범죄예측 방식에 지능형 CCTV 영상분석 기술을 더한 것이다. 또한 한국전자통신연구원은 사람 재식별기술(Person Re-ID)을 활용하여, 전자발찌 착용자처럼 고위험군 특정인의 경로를 분석하고 인근 CCTV로 사람을 찾을 수 있도록 지원해 관리자 분석 및 위험 행동 징후를 신속히 파악한 뒤 대응을 할 수 있는 '성범죄 전과자 관리 기술'도 개발할 계획이다.[29]

5. 범죄예측 시스템의 문제점

1) 헌법과 법률상의 권리침해

개인정보를 개인식별정보이든 비(非)개인식별정보이든 민감정보이든 비(非)민감정보이든 불문하고 범죄예측 시스템으로 분석하게 되면 필연적으로 헌법상 보장된 국민의 기본권인 사생활의 비밀과 자유 및 개인정보자기결정권 등이 침해될 수밖에 없다. 왜냐하면 아무리 익명화되고 비식별조치가 된 정보라도 다수의 정보가 모여지게 되면 특정 개인을 추정할 수 있게 되는데 수많은 정보가 모여지는 빅데이터의 특성상 개인의 프라이버시와 개인정보자기결정권이 침해될 수 밖에 없다. 실제로 2018년 서울시 민생사법경찰단은 불법대부, 다단계 등 민생범죄 수사에 범죄예측 시스템을 이용한 '인공지능 수사관'활용해 수사를 진행하는 것이 개인정보보호법을 위반했다는 개인정보보호위원회의 결정이 나오면서 수사에 제동이 걸렸다.[30]

헌법상 평등권에 대한 침해도 우려된다. 즉, 범죄예측 시스템으로 범죄의 조사대상이 된 특정인과 그렇지 않은 일반인 사이에는 형평성의 문제가 발생한다. 정보의 주체인 국민에 대해서 기본권 보호가 균등하게 되어야 하는데 조사 대상자는 그렇지 않은 일반인과 달리 기본권이 더 제한된다는 것이다. 또한 범죄예측 시스템이 과거의 범죄와 연관된 사건이나 전과 등의 데이터를 바탕으로 앞으로 발생할 범죄를 예측하기 때문에 범죄경력이 과거에 있었던 특정인을 일반인보다 더 집중적으로 감시 및 통제하게 된다. 이러한 감시 및 통제는 대상자를 잠재적인 범죄자로 낙인을 찍는 효과를 유발하게 되는데 결과적으로 대상자를 일반인과 차별하게 되는 문제가 발생한다. 또한 조사 대상자의 범위를 어디까지로 설정할지도 문제가 된다. 만약 대상자를 피의자, 피고인, 수형자 등으로 한정할 경우 형의집행및수용자처우에관한법률과 개인정보보호법 등에 규정이 있는지도 문제가 될 것이다.[31]

과도하게 범죄예측 시스템을 사용하여 범죄의 예방을 시도하는 것은 일반인의 행복추구권과 행

---

29) 뉴시스, "전자통신연구원, AI와 CCTV로 범죄 예측한다", 2020.01.02.
<https://newsis.com/view/?id=NISX20200102_0000877333>

30) 서울시 민생사법경찰단(민사단)의 '인공지능을 활용한 민생범죄 수사지원 분석사업'(인공지능 수사관)이 개인정보보호법을 위반했다는 개인정보보호위원회(개보위)의 결정이 나왔다. 빅데이터·인공지능(AI) 등 새로운 기술을 범죄수사 등에 사용할 경우 정보인권을 고려한 제도적 환경이 조성돼야 한다는 지적이 나온다. 한겨레, "'AI 수사관, 개인정보 수집은 위법"…SNS 불심검문 제동 걸려", 2019. 6.11.
http://www.hani.co.kr/arti/economy/it/897375.html

31) 윤해성 등, 앞의 글, 141-146면.

동 자율권을 침해하게 될 것이다. 따라서 범죄예측 시스템의 사용에 의한 부작용도 고려하여야 한다. 예를 들어 범죄예측 시스템이 특정인 갑이 12시간 이내에 범죄를 저지를 것이라는 예측결과를 도출했다고 하자. 그래서 담당자가 적절한 감시와 통제에 나섰고 그 결과 범죄가 방지되었다면, 그것이 그 담당자의 감시와 통제로 인하여 범죄가 예방된 것인지 아니면 애초부터 범죄예측 시스템의 예측 자체가 틀린 것인지 누구도 알 수가 없다. 실제로 오류가 발생한 때도 오류가 발생하였다는 사실을 알 수 없으므로 오류를 바로잡으려는 기회조차 가질 수 없다는데 더 문제가 있다. 오류를 검증할 수 있는 유일한 방법은 위와 같은 사안에서 담당자가 감시와 통제에 나서지 않는 것이다. 만약 12시간 이내에 갑이 정말로 범죄를 저지른다면 범죄예측 시스템이 올바르게 예측한 것이며, 그렇지 않다면 적어도 당해 예측에 대해서는 범죄예측 시스템이 틀린 것이다. 그러나 이와 같은 무작위성에 의존하여 법집행을 할 수는 없다. 범죄예측 시스템이 범죄 발생을 예측하였다면 담당자는 반드시 감시와 통제에 나서야 한다. 그런데 범죄예측 시스템의 오류율이 1%라고 하더라도 100회의 감시와 통제 시도 중에 1회는 아무런 잘못이 없는 사람에게 불필요한 감시와 통제를 한 결과가 될 것이다. 그런데 그것이 옳지 않은 감시와 통제였다는 것은 알 수 없으므로 억울하게 부당한 대우를 받은 사람은 사후에 구제를 받을 가능성도 희박하다.[32] 범죄예측의 정확성이 보장되지 않으면 부정확한 범죄예측을 근거한 예방책은 유해하거나 무익하다. 아직 벌어지지 않은 사건을 다루는 범죄예측 시스템은 고도의 정확성이 보장되어야 한다.[33] 부정확한 범죄예측으로 개인을 감시하고 통제한다면 개인의 자율성과 인권을 크게 침해하게 될 것이다.[34]

2) 알고리즘의 불투명성, 편향성 및 부정확성

인공지능에는 주어진 정보(data)를 자동적으로 분석하고 처리하는 과정이 있는데 이것을 인공지능의 알고리즘이라고 한다.[35] 그런데 이러한 알고리즘이 정보를 입력받아 결론을 도출되는 과정이 불투명하여 충분히 공개되거나 설명되지 않는다는 문제점이 알고리즘의 불투명성이다.[36] 이러한 알고리즘의 불투명성은 범죄예측 시스템의 대표적인 문제점 중의 하나이다. 왜냐하면 범죄예측 시스템으로 인하여 부정확한 범죄예측이 발생하게 되면 왜 잘못된 범죄예측이 발생했는지 원인을 규명하여야 하는데 인공지능의 처리과정 자체가 공개되지 않거나 설명되지 않아서 원인의 규명은 물론 피해나 그 책임의 소재를 찾는데 어려움을 준다는 것이다.[37]

알고리즘이 불투명한 원인은 소스코드와 같은 알고리즘의 세부 정보가 공개되더라도 전문적 지식이 없는 일반인은 이를 읽고 이해하는 것은 매우 어렵다는 점이다. 또한 알고리즘을 개발하거나 설계한 전문가들조차도 인공지능의 작동원리나 내부구조를 정확하게 파악할 수 없다는 알고

32) 윤영석, 앞의 글, 169면.
33) Mark C. Niles, "Preempting Justice : "Precrime" in fiction and in fact", Seattle journal for social justice Vol. 9 Iss. 1, 2010. 1., p. 303.
34) Jennifer C. Daskal, "Pre-crime restraints: the explosion of targeted, noncustodial prevention", Cornell law review Vol. 99 Iss. 2, 2014. 1, p. 332.
35) 크리스토퍼 스타이너/박지유(역),『알고리즘으로 세상을 지배하라』, 에이콘(2016). 14면.
36) Andrea Roth, Trial by Machine, 104 Geo. L. J. 1245 (2016), 제1269쪽.
37) 양종모, "인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안", 법조 제723권, 법조협회, 2017. 80면.

리즘의 특성을 ‘블랙박스(black box)’라고 하는데 이것은 알고리즘의 불투명성의 원인으로 설명하고 있다.[38] 그리고 알고리즘 자체가 특정 개인이나 기업의 소유물로서 영업비밀 등으로 보호되고 있다. 영업비밀과 같이 개인이나 기업이 보유한 기술상 또는 경영상 정보는 기업의 경쟁력을 좌우하는 중요한 요인이기 때문에 명확한 공개를 요구하기가 쉽지 않다.[39]

범죄예측의 결과가 특정 계층이나 소수집단에 불리하도록 편향되게 산출되는데 이 때문에 범죄예측이 공정하지 못하다는 문제가 알고리즘의 편향성이다. 알고리즘의 편향성은 편향된 데이터의 문제일 수도 있지만, 알고리즘의 개발자가 의도적으로 알고리즘을 편향적으로 설계할 때도 발생할 수 있다. 또한 개발자가 지닌 편향된 지시과 경험이 무의식중에 알고리즘에 그대로 투영되어 발생할 수도 있다.[40] 그러나 알고리즘의 편향성의 가장 주요한 원인은 편향된 데이터이다. 알고리즘이 학습하는 데이터들 중에는 인종, 성별 등 차별적인 요인들이 많은데 그 결과 편향적인 판단을 하게 되는 것이다. 예를 들어 과거 유죄판결과 같이 범죄사건을 토대로 한 데이터가 많은데 이러한 데이터를 기초로 미래의 범죄를 예측하다 보니 과거 범죄사건과 관련된 지역이나 사람은 그렇지 않은 지역이나 사람보다 더 집중적인 감시와 통제를 받아 우범지역이나 범죄자로 낙인찍히게 된다는 것이다. 이러한 편향성의 대표적인 사례로 Northpoint가 개발한 교정 범죄자 프로파일링 관리시스템인 컴파스(COMPAS, Correctional Offender Management Profiling for Alternative Sanctions)는 범죄자 재범률을 계산하여 재범의 위험성을 예측한 결과 흑인 피고인을 고위험군으로 잘못 분류하는 편향이 2배나 되는 오류가 나타났다. 즉 흑인이 강력범죄를 저지를 위험성이 백인보다 77% 더 높게 측정하여 인종차별적이란 비판을 받았다.[41]

범죄예측 시스템은 축적된 데이터를 바탕으로 장래 범죄의 발생을 예측하게 되는데, 이러한 범죄예측은 범죄 발생 가능성이 높다는 확률을 근거로 하는 것이다. 따라서 100% 정확하게 예측하는 것은 현재로서는 불가능하다. 결과적으로 부정확한 예측이 발생할 수 있다. 이러한 알고리즘의 부정확성 또한 문제가 된다. 예를 들어 부정확한 예측으로 인하여 범죄를 저지를 것으로 예측하고 예방조치를 취했지만 실제로는 범죄를 저지르지 않은 잘못된 긍정 오류(false positive)의 경우 무고한 특정인을 혐의자로 만들게 되며, 부당하게 혐의를 쓴 특정인은 사회적 비난과 감시 등으로 인권이 침해되는 대가를 치르게 된다. 반대로 더 이상 범죄를 저지르지 않을 것으로 예측된 특정인에게 예방조치 등을 하지 않았지만 그 특정인이 실제로 범죄를 저지르게 된 잘못된 부정 오류(false negative)의 경우 범죄를 방치하는 결과가 되는데, 이 역시 사회방위에 위협이 되고 범죄로 인한 피해로 인해 사회전체의 비용이 증가하게 된다.

결국 인공지능 알고리즘의 불투명성, 편향성 및 부정확성이 해결되지 아니하면 예측의 결과 및 판단의 신뢰성을 담보하기는 어렵다. 이는 최종적으로 범죄예측 시스템에 대한 불신을 넘어 법집행을 포함한 사법절차 전체에 불복을 초래한다.[42]

3) 새로운 통제수단으로서 범죄예측 시스템

38) 양종모, 앞의 글, 81면, 93면.
39) State v. Loomis, 881 N. W. 2d 749 (Wis. 2016).
40) 양종모, 앞의 논문, 74-77면.
41) Amy B. Cyphert, Reprogramming Recividism: The First Step Act and Algorithmic Prediction of Rick, 51 Seton HALL L. REV. 331(2020), p.339.
42) 윤해성 등, 앞의 글, 254, 255쪽.

범죄예방이라는 명목으로 AI 범죄예측 시스템은 발생하지도 않은 사건에 대한 구속과 처벌을 가능하게 할 수 있다. 독재자나 권력자들은 범죄예측 시스템의 설계자나 개발자를 조종하여 빅데이터를 가지고 편향된 알고리즘을 만들 수 있다. 독재자나 권력자들은 이것을 가지고 자신들의 반대세력이나 정적들을 고위험군의 범죄 용의자 그룹으로 설정해 놓고 감시와 통제를 손쉽게 하려고 할 것이다. 그러나 감시나 통제대상이 된 일반인들은 블랙박스와 같은 알고리즘의 불투명성 때문에 그러한 사실을 알지도 못할 것이다. 범죄예측 시스템이 사회통제 시스템으로 둔갑하면 얼마든지 '빅브라더'의 출현은 얼마든지 가능하게 된다. 소수의 권력층이 정보를 독점한 사회에서는 범죄예측 시스템은 고위험군으로 분류된 사람들에 대한 감시와 감독 등을 정당화하는 새로운 통제수단으로 등극할 것이다.[43]

Ⅲ. 현행 구속기준의 문제점

1. 구속사유의 추상성

1) 형사소송법 제70조 1항의 구속사유

형사소송법 제70조 1항은 죄를 범하였다고 의심할 만한 상당한 이유가 있고 ① 피의자가 일정한 주거가 없는 때, ② 피의자가 증거를 인멸할 염려가 있는 때, ③ 도망 또는 도망할 염려가 있는 때를 구속사유로 규정하고 있다. 현행 구속사유의 개념은 그 자체가 추상적이다. 이러한 추상적인 구속사유만으로는 특정 피의자에 대하여 과연 구속영장이 발부될 것인지를 법률전문가들조차 예측하는 것이 결코 쉽지 않다. 그러므로 구체적 사건에 있어서 관련 당사자나 일반국민들이 구속영장이 발부되는 경우와 기각되는 기준이 무엇인지 알기 어려운 경우가 대부분이다. 또한 형사소송법 제70조 1항의 구속사유 즉 주거부정, 증거인멸 및 도망의 염려만을 기준으로 삼을 경우, 안정된 주거와 소득이 많고 사회적 지위가 높을수록 구속가능성이 낮고, 일정한 주거·소득이 없거나 사회적 지위가 낮을수록 구속될 가능성이 상대적으로 커 '유전불구속 무전구속'의 논란이 제기될 수 있다.

2) 제70조 2항의 구속고려요인 또는 구속요인

2007년 형사소송법 일부개정을 통해서 "법원은 제1항의 구속사유를 심사함에 있어서 범죄의 중대성, 재범의 위험성, 피해자 및 중요 참고인 등에 대한 위해우려 등을 고려해야 한다"라고 제70조 2항을 신설하여 구속사유 판단시 범죄와 관련된 중요한 형사정책적 요인을 의무적으로 고려

43) 유강하, "빅데이터와 사물인터넷 시대의 비판적 해석과 인문학적 상상력-영화<마이너리티 리포트>를중심으로", 시민인문학 제30호, 2016, 100면 이하.

하여야 함을 명시하고 있다(이처럼 구속사유 판단에 고려하여야 할 요인을 구속고려요인 또는 구속요인이라 한다. 이하에서는 '구속요인'으로 함). 이렇게 개정하게 된 이유는 "종래의 수사 및 재판실무 관행에서도 구속여부를 판단할 때 사안의 중대성이나 재범의 위험성 등을 고려하여 왔으나,[44] 이러한 실무 관행에 대해서는 명문 규정이 없고, 개정 전 법의 구속사유만으로는 구속여부 판단의 기준이 구체화되지 아니하여 결과적으로 검사나 법관의 해석재량의 폭이 지나치게 넓어지며 결과적으로 구속여부에 대한 예측가능성이 없기 때문에 국민의 불신을 초래할 우려가 있다는 지적이 제기 되었다. 이에 따라 지금까지의 실무관행을 형사소송법에 반영함으로써 선진각국에서 인정하는 예방적 구금과 유사한 개념을 도입하고 법규정과 실무 사이의 괴리를 해소하여 구속제도의 운영과 관련된 국민들의 신뢰를 회복하고자 하는 것"[45]이라고 하면서 범죄의 중대성, 재범의 위험성, 피해자 및 중요 참고인 등에 대한 위해우려 등을 단순 고려사항이 아니라 구속사유에 준할 정도의 사유라고 이해해야 한다고 주장한다.[46] 그러나 이에 반해서 위 규정의 의의를 축소해석하는 경우에는 위 사유들을 단순히 구속사유를 심사함에 있어서 고려해야 할 사항의 하나일 뿐이라고 보는 견해도 있다.[47]

그러나 어느 견해도 범죄의 중대성, 재범의 위험성, 피해자 및 중요 참고인 등에 대한 위해우려 등의 구속요인을 어떻게 얼마나 어떤 방법으로 고려하여야 하는지 구체적으로 명시하고 있지 않다. 어느 정도가 중대한 범죄인지, 재범의 위험성은 어떠한 경우에 인정할 것인지, 참고인의 범위와 위해의 우려는 어느 정도까지 인정할 것인지에 대한 어떠한 기준제시가 없다는 것이다. 이들 구속요인도 역시 법원의 판단 여하에 맡겨져 있을 뿐 국민에게 어떤 구체적인 기준의 제시도 없다. 또한 70조 2항은 "고려하여야 한다"고 규정되어 있어 '고려' 자체는 법원의 의무사항에 해당한다. 따라서 영장발부여부를 담당하는 법관으로서는 위 고려요인들을 필수적으로 고려하여야 할 것인데, 실제 고려하였는지 여부에 관하여는 외부에서 전혀 알 수도 없다는데 문제의 심각성

---

44) 종전까지 구속실무에서는 재범의 위험성이 있는 중대한 범죄에 대해서는 도망의 염려가 있는 것으로 추정해 왔기 때문이다. 조성용·신태용, "예방구속 도입의 정당성 여부와 입법론적 대안", 저스티스, 통권 제109호, 한국법학원, 2009, 234면.

45) "외국의 예방적 구금 개념과 유사한 범죄의 중대성, 재범의 위험성, 피해자·중요 참고인 등에 대한 위해 우려 등이 구속사유 심사의 고려사항으로 자리잡게 되었다. … 첫째, '범죄의 중대성'은 필요적 보석의 제외사유인 '사형, 무기 또는 장기 10년이 넘는 징역이나 금고에 해당하는 죄를 범하였다고 의심할 만한 상당한 이유가 있는 때'를 기초로 성안된 요건으로 중대 범죄를 저지른 자에게는 높은 처단형이 예상되는 만큼 주로 구속사유 중 '도주의 우려'를 판단할 때 적극적인 요인으로 고려된다. 둘째, '재범의 위험성'은 피고인의 전과에 의하여 당해 범죄가 누범 또는 상습범에 해당하거나 범죄의 특성상 반복 범죄의 개연성이 높을 것을 의미하는 것으로 이해되며, 역시 재범의 위험성이 높은 자에 대해서는 높은 처단형이 예상되므로 주로 구속사유 중 '도주의 우려'를 판단할 때 적극적인 요인으로 고려된다. 셋째, '피해자·중요 참고인 등에 대한 위해 우려'는 피해자·중요 참고인 등에 대한 위해와 증거인멸을 방지하려는 것에서 입법 이유를 찾아볼 수 있으므로 주로 구속사유 중 '증거인멸의 우려'를 판단하는 주요 기준으로 기능할 것"이라고 주장하여 긍정적으로 평가하고 있다. 법무부, 개정형사소송법, 2007, 40면 이하.

46) 이경재, "개정 형사소송법상 인신구속제도에 관한 고찰", 형사법연구 제19권 제4호(통권 제33호, 2007년 겨울호), 한국형사법학회, 9면.

47) 2007년 형사소송법 개정과정에 관하여, "당초 법개정 과정에서 일부 법사위원들은 '사안의 중대성'을 구속사유에 포함시켜야 한다는 입장을 보였으나, 구속을 예외적인 강제처분으로 보고 있는 헌법과 형소법 태도에 어긋나고 헌법상 무죄추정의 원칙과 충돌할 우려가 있다는 지적 등을 감안해 별도의 구속사유로 인정하지 않는 대신 필요적 고려사항으로 규정했다."고 주장("구속영장 기각률 이전보다 높아졌다", 2008.5.1.자 인터넷 법률신문 ; 법원행정처, 형사소송법 개정법률 해설, 2007, 1면 이하 참조)하거나, "새로 요건으로 추가된 범죄의 중대성, 재범의 위험성, 피해자 및 중요 참고인 등에 대한 위해 우려는 독립적인 구속영장 발부요건이 아니라 구속영장 발부 심사시 고려할 사항에 불과하다."고 주장("피고인 구속 '실형기준원칙' 절반의 실패", 2009.4.23.자 인터넷 법률신문 중 당시 서울중앙지방법원 김도형 영장전담판사 발언부분 참조)하여 그 의미를 제한적으로 파악하는 견해가 있다.

이 있다.[48]

결국, 개정 형사소송법이 제70조 2항을 추가했더라도 국민들의 구속여부에 대한 예측가능성을 개선하지는 못했다.

2. 개별기준으로 인한 업무의 비효율과 국민의 혼란

형사소송법상의 구속사유가 그 추상성으로 말미암아 객관적 기준으로 사용되기에는 부족함이 있어 법원은 '인신구속사무의 처리에 관한 예규(재형 2003-4)'[49]을 통해서 그리고 검찰은 '구속수사기준에 관한 지침(대검예규 기획 제400호)'[50]을 통해서 각각 개별 구속기준을 가지고 있다

48) 실제 구속영장발부여부를 검토하는 법관으로서는 일한 사정들을 고려하고 판단하였을 것으로 생각된다. 다만, 이 점은 단순히 법관의 머릿속에서만 존재하는 것일 뿐이므로, 실제 고려하였는지, 어느 부분을 얼마만큼 어떻게 고려하였는지도 정확히 알 수 없는 것이 현실이다.

49) 인신구속사무의 처리에 관한 예규(재형 2003-4), 재판예규 제900호 제정 2003. 9. 3.

제52조(증거를 인멸할 염려) 구속의 사유 중 증거를 인멸할 염려는 다음 각호의 요인을 종합적으로 고려하여 판단한다.

1. 인멸의 대상이 되는 증거가 존재하는지 여부
2. 그 증거가 범죄사실의 입증에 결정적으로 영향을 주는지 여부
3. 피의자측에 의하여 그 증거를 인멸하는 것이 물리적·사회적으로 가능한지 여부
4. 피의자측이 피해자 등 증인에 대하여 어느 정도의 압력이나 영향력을 행사할 수 있는지 여부

제53조 (도망할 염려) 구속의 사유 중 도망할 염려는 다음 각호의 요인을 종합적으로 고려하여 판단한다.

1. 범죄사실에 관한 사정
 (1) 범죄의 경중, 태양, 동기, 횟수, 수법, 규모, 결과 등 (2) 자수 여부
2. 피의자의 개인적 사정
 (1) 직업이 쉽게 포기할 수 있는 것인지 여부 (2) 경력, 범죄경력, 범죄에 의존하지 아니하고도 생계를 유지하였는지 등 그 동안의 생계수단의 변천 (3) 약물복용이나 음주의 경력, 피의자의 도망을 억제할 만한 치료 중인 질병이 있는지 또는 출산을 앞두고 있는지 여부 (4) 다른 곳 특히 외국과의 연결점이 있는지 여부, 여권의 소지 여부 및 여행 특히 해외여행의 빈도
3. 피의자의 가족관계
 (1) 가족간의 결속력 (2) 가족 중에 보호자가 있는지 여부 (3) 배우자 또는 나이가 어리거나 학생인 자녀가 있는지 여부 (4) 연로한 부모와 함께 거주하거나 부모를 부양하고 있는지 여부 (5) 피의자에 대한 가족들의 의존 정도 (6) 가족들이 피의자에게 양심에 호소하는 권고나 충고를 하여 피의자를 선행으로 이끌만한 능력과 의사가 있는지 여부
4. 피의자의 사회적 환경
 (1) 피의자의 지역사회에서의 거주기간 및 지역사회에서의 정착성의 정도 (2) 피의자 가족의 지역사회와의 유대의 정도 (3) 교우 등 지원자가 있는지 여부

50) 구속수사기준에 관한 지침, 대검예규 기획 제400호, 2006. 6. 12.

제6조(증거인멸의 염려) ① 피의자가 증거를 인멸할 염려가 있는지 여부를 판단할 때에는 다음 각 호의 요인을 고려하여야 한다.

1. 사안의 성격,
2. 사안의 경중,
3. 증거의 수집정도,
4. 피의자의 성행, 지능과 환경,
5. 물적 증거의 존재 여부와 현재 상태,
6. 공범의 존재 여부와 현재 상태,
7. 피해자, 참고인 등 사건 관계인과 피의자와의 관계,
8. 수사 협조 등 범행 후의 정황9. 범죄 전력

제7조(도망 또는 도망할 염려) ② 피의자가 도망할 염려가 있는지 여부를 판단할 대에는 다음 각 호의 요인을 고려하여야 한다.

1. 사안의 경중
2. 범행의 동기, 수단과 결과
3. 전문적·영업적 범죄여부

(이들 구속기준에는 구속사유를 판단하는 구속요인들을 열거하고 있다). 각 기관들이 개별기준을 가지는 것은 각 당사자의 형사소송 절차에 있어서의 역할에 비추어 당연한 일로 받아들여져 왔지만, 구속영장실질심사제도가 전면적으로 시행되는 이상 이는 불필요한 사법행정적 업무를 발생시키는 한 요인이라고 할 수 있다. 이것은 한정된 인적·물적 자원의 효율적 활용을 저해하는 요인이 되고 있다. 구속업무에 구속영장의 기각률이 높아지고 있는데 법원과 검찰간에 영장발부기준과 관련하여 의견이 일치되지 않아 불필요한 인력과 시간의 낭비요인이 되고 있는 것이다.[51] 법원과 검찰간에 구속기준이 공유되었다면 검찰은 구속영장이 기각될 것이 예상되는 사안에서는 신청을 하지 않음으로써 불필요한 인력과 시간의 낭비를 막을 수 있을 것이고 인권보호에도 기여할 수 있을 것이다.

그리고 국민들의 입장에서도 현행법상 구속사유의 추상성 때문에 구체적 판단이 어려운 현실에서 각 기관마다 다른 기준을 적용하여 더 혼란을 가중시키고 있다. 또한 법원과 검찰의 개별 기준이라는 것이 일반 국민들이 접근하기 어려운 예규나 지침의 형식으로 존재하기 때문에 국민들에게는 더더욱 수용하기가 힘들다. 물론 형사소송법상에 구속사유에 관한 모든 것을 명시할 수 없는 입법기술상 한계가 있을 수 있다는 것은 인정할 수 있으나, 국민의 중요한 신체의 자유를 제한하는 구속사유를 판단하는데 있어 일반 국민들이 접근하기 어려운 예규 등의 형식으로 존재하게 한다는 것은 죄형법정주의와 국민의 법적 생활의 안정성과 예측가능성을 위해서도 지양되어야 한다.

3. 구속요인의 동등 취급의 문제와 평가방법의 부재

'주거부정', '증거인멸의 우려', '도망의 위험'의 구속사유가 존재하는지 여부를 판단하는데 고려하여야 할 구속요인들에는 제70조 2항에서 열거하고 있는 '범죄의 중대성', '재범의 위험성', '피해자 및 중요 참고인 등에 대한 위해우려' 등이 있다. 그리고 법원과 검찰의 개별 기준인 '인신구속사무의 처리에 관한 예규(재형 2003-4)'와 '구속수사기준에 관한 지침(대검예규 기획 제400호)에도 구속요인들을 열거하고 있는데 아래의 표에서 정리하면 다음과 같다.

【개별구속요인】

4. 피의자의 성행, 연령, 건강 및 가족관계
5. 피의자의 직업, 재산, 교우, 조직·지역 사회정착성, 사회적 환경
6. 주거의 형태 및 안정성
7. 국외 근거지의 존재 여부, 출국 행태 및 가능성
8. 수사 협조 등 범행 후의 정황
9. 범죄 전력
10. 자수 여부
11. 피해자와의 관계, 피해 회복 및 합의 여부

51) 법원의 "인신구속사무의 처리에 관한 예규"와 검찰의 "체포·구속 업무처리지침" 등 이들 기준을 적용한 결과를 비교해 보면 양 기관들의 일치율에서 10% 이상 차이가 나며 이는 법원과 검찰의 인신구속에 대한 인식에 있어서 간과할 수 없는 차이가 있음을 징표한다고 할 수 있을 것이다. 한국형사정책연구원, 21세기 형사사법개혁의 방향과 대국민 법률서비스 개선방안(Ⅳ) -수사절차상 인권보호와 효율성제고-, 2004, 274-277면.

<table>
<tr><td colspan="2" rowspan="2">구분</td><td colspan="3">개별 구속(고려)요인</td></tr>
<tr><td colspan="2">재판예규</td><td>대검예규</td></tr>
<tr><td rowspan="5">구속사유판단기준</td><td>증거를 인멸할 염려</td><td colspan="2">① 인멸의 대상이 되는 증거가 존재 여부
② 사건에 결정적 증거인지 여부
③ 인멸의 물리적·사회적 가능성
④ 피의자측이 피해자 등 증인에 대하여 어느 정도의 압력이나 영향력을 행사할 수 있는지 여부 등 검토</td><td>① 사안의 성격
② 사안의 경중
③ 증거의 수집정도
④ 피의자의 성행, 지능과 환경
⑤ 물적 증거의 존재 여부와 현재 상태
⑥ 공범의 존재 여부와 현재 상태
⑦ 피해자, 참고인 등 사건 관계인과 피의자와의 관계
⑧ 수사 협조 등 범행 후의 정황
⑨ 범죄 전력</td></tr>
<tr><td rowspan="4">도망하거나 도망할 염려</td><td>범죄사실 관련 사정</td><td>① 범죄의 경중, 태양, 동기, 횟수, 수법, 피해규모, 결과 등
② 자수 여부(자수 경위)</td><td rowspan="4">1. 사안의 경중
2. 범행의 동기, 수단과 결과
3. 전문적·영업적 범죄여부
4. 피의자의 성행, 연령, 건강 및 가족관계
5. 피의자의 직업, 재산, 교우, 조직·지역 사회정착성, 사회적 환경
6. 주거의 형태 및 안정성
7. 국외 근거지의 존재 여부, 출국 행태 및 가능성
8. 수사 협조 등 범행 후의 정황
9. 범죄 전력
10. 자수 여부
11. 피해자와의 관계, 피해 회복 및 합의 여부</td></tr>
<tr><td>개인사정</td><td>① 직업이 쉽게 포기할 수 있는 인지 여부(그 동안의 생계 수단 변천 등), ② 경력, 범죄경력, ③ 약물 복용 경력, 질병 치료, 여성의 경우 출산을 앞두고 있는지 여부, ④ 외국과의 연결점이 있는지, 해외여행의 빈도 등</td></tr>
<tr><td>가족관계</td><td>① 가족간의 결속력 ② 가족 중에 보호자가 있는지 여부 ③ 배우자 또는 나이가 어리거나 학생인 자녀가 있는지 여부 ④ 연로한 부모와 함께 거주하거나 부모를 부양하고 있는지 검토 ⑤ 피의자에 대한 가족들의 의존정도 ⑥ 가족들이 피의자에게 양심에 호소하는 권고나 충고를 하여 피의자를 선행으로 이끌 만한 능력과 의사가 있는지 여부</td></tr>
<tr><td>사회환경</td><td>① 피의자의 지역사회에서의 거주기간 및 지역사회에서의 정착성의 정도 ② 피의자 가족의 지역사회와의 유대의 정도 ③ 교우 등 지원자가 있는지 여부</td></tr>
</table>

그런데 문제는 검찰과 법원이 구속사유를 판단하는데 고려하는 개별 구속요인들을 단순히 나열만 하고 있다는 것이다. 범죄의 유형에 따라서는 다른 구속요인보다 상대적으로 더 중요하게 취급되어야 할 구속요인이 있을 수도 있는데, 각 요인별 특성을 고려하지 않고 모두 동등한 비중으로 고려되도록 구속기준이 설정된 것이 문제이다. 또한 주관적 판단여지가 많은 요인에 그 평가방법을 구체화하지 않고 단순히 고려만 하도록 한 것도 문제이다.[52] 즉, 구속요인들을 어떻게 얼마나 어떤 방법으로 고려하여야 하는지 구체적으로 명시하고 있지 않다. 어느 정도가 중대한 범죄인지, 재범의 위험성은 어떠한 경우에 인정할 것인지, 참고인의 범위와 위해의 우려는 어느 정

52) 예를 들면, 도망의 위험을 결정하는 구속요인인 피의자의 가족관계를 판단할 때 도망의 위험성을 상, 중, 하로 판단한다든지 얼마든지 주관적 판단여지가 많은데 그 평가방법을 구체화하지 않고 단순히 고려만 하도록 한 것은 비판의 여지가 있다. 김병수, "구속기준의 객관화와 구속절차의 투명화", 외법논집 제34권 제1호, 2010, 189면.

도까지 인정할 것인지에 대한 어떠한 기준제시가 없다는 것이다. 이들 구속요인도 역시 법원의 판단 여하에 맡겨져 있을 뿐 국민에게 어떤 구체적인 기준의 제시도 없다.[53]

4. 불복방법의 부재

현재 학계의 주류적 견해[54]와 대법원 판례[55]가 영장결정에 대한 항고나 재항고를 통한 불복방법을 인정하지 않는다. 이에 대하여 '영장항고제도'를 도입하여야 한다는 주장이 있다.[56] 영장재판도 그 판단의 적법타당성을 보장하기 위해서는 사후적 심사의 가능성이 원칙적으로 열려 있어야 한다는 점[57]과 영장담당판사가 절차규정을 준수하지 않았거나 그 재량을 남용하여 자의적 판단을 하는 경우를 시정하여야 한다는 점,[58] 영장결정에 대한 항고를 허용하지 않게 되면 구속사유에 관한 선례가 축적될 수 없어 영장재판의 형평성과 예측가능성을 확보할 수 없다는 점[59] 등을 들어 현재의 대법원의 태도를 비판하는 견해[60]도 유력하게 주장되고 있다. 하지만 영장항고제도라는 것도 결국 법원의 재판을 통해서 영장발부여부를 판단하는 것이기 때문에 법원이 구체적인 판단을 하지 않고 영장결정처럼 불명확한 판단으로 일관한다면 영장항고제도의 도입취지가 무색하게 될 우려가 있다. 이것은 법원의 판단을 요구하는 영장항고제도가 가지는 본질적인 취약점이 될 수도 있다. 그러므로 영장항고제도에서도 중요한 것은 법원이 영장결정의 당부를 판단할 때 개별 구속요인들을 어떻게 평가할 것인가이다. 결국 영장항고제도도 구속사유를 어떻게 평가할 것인가가 문제로 귀결된다.

Ⅳ. AI를 활용한 구속기준의 재정립

1. AI를 활용한 구속기준 정립의 필요성

53) 김병수, "구속요인의 계량화를 통한 구속기준 정립", 형사법연구 제24권 2호, 2012, 275면.

54) 배종대 외. 신형사소송법, 홍문사, 2011, 141면; 손동권, 형사소송법, 세창출판사, 2010, 264면; 송광섭, 형사소송법, 유스티아누스, 2003, 369면; 신양균, 신판 형사소송법, 법문사, 2009, 182면; 이은모, 형사소송법, 박영사, 2010, 242면; 이재상, 신형사소송법, 2011, 251면; 임동규, 형사소송법, 법문사, 2009, 192면; 정웅석, 형사소송법, 대명출판사, 2007, 234면.

55) 법원의 결정에 항고할 수 있다고 규정한 제402조가 말하는 법원은 수소법원만을 의미하며, 영장결정을 하는 지방법원 판사는 수소법원이 아니기 때문에 이에 해당하지 않으며, 준항고에 관한 제416조는 수소법원의 구성원으로서의 재판장이나 수명법관의 재판에 관한 것인데 지방법원판사는 이에도 해당하지 않는다. 대법원 2006.12.18. 선고 2006모646 결정; 대법원 2005.3.31. 선고 2004모517 결정; 대법원 1986. 7.12. 선고 86모25결정; 대법원 1958. 3.14. 선고 4290형항9결정.

56) 김경수, "영장항고제도 도입 필요성에 관한 연구", 형사정책연구, 제84호, 한국형사정책연구원, 2010.; 변종필, "구속제도의 이론과 실제", 형사정책연구, 제82호, 한국형사정책연구원, 2010; 신의기, "검찰과 법원의 구속영장신청 및 발부기준 차이와 해결방안", 형사법의 신동향, 통권 제26호, 대검찰청 미래기획단, 2010. 6; 강동범, "구속영장재판에 대한 불복 가부", 형사법연구 제21권 3호, 한국형사법학회, 2009. 등

57) 김재봉, "개정 형사소송법과 수임판사의 재판에 대한 불복", 형사법연구 제19권 제4호, 한국형사법학회, 2007. 49면.

58) 김재봉, 앞의 논문, 49면.

59) 황정근, "구속영장실질심사제도의 개선방안", 형사정책연구 제8권 제4호, 1997, 58면.

60) 박병규, "영장재판에 대한 항고", 법조, 법조협회, 2007.12, 387면.

‘증거인멸’, ‘도주우려’ 등 현행 구속사유(형사소송법 제70조)가 너무 추상적이어서 구체적 사건에서 정확한 구속기준이 뭔지 알기도 어렵고 예측하기도 어렵기 때문이다. 게다가 법률전문가인 검찰과 법원도 국민들이 공감할 수 있는 명확한 판단기준을 제시하고 못하고 있다. 이 때문에 국민들은 어떤 경우에 구속이 되는지를 몰라 인신구속의 두려움을 해소하기 위하여 비싼 수임료를 부담해서라도 전관출신의 변호사를 찾고 있는 것이 현실이며, 재벌총수와 정치인 등 돈 있는 사람과 힘없고 돈 없는 서민의 구속기준이 다르다고 생각하여 ‘유전불구속 무전구속’이란 말까지 하여 왔다. 국민들의 법률생활을 불안정하게 하고 예측가능성을 곤란하게 하며 형평성에 의구심까지 갖게 하는 구속기준을 재정립하는 것은 시대의 요청인 것이다.

과거에는 구속판단은 사실상 법관의 감(感)에 의한 주관적 판단에 의존하였던 것이 사실이다. 그러나 앞으로의 구속판단은 객관적인 데이터에 근거한 객관적인 판단이어야 할 것이다. 그러기 위해서는 수사자료, 조서, 영장, 공소장, 판결문 등의 같이 형사사법기관들이 형사사법업무와 관련하여 축적한 형사사법정보를 바탕으로 객관적인 판단을 하여야 할 것이다. 이러한 정보들은 방대하여 빅데이터에 속한다. 오늘날 이러한 빅데이터는 인간이 수집, 분석 및 처리하는데 한계가 있으므로 머신러닝 및 딥러닝 기술에 근거한 인공지능이 수집, 분석 및 처리를 하고 있다. 이처럼 방대한 사법정보를 인공지능이 처리하여 구속기준을 재정립할 수 있다면 법관의 감에 의한 주관적인 구속판단에서 데이터에 기반한 객관적인 구속판단으로 구속판단의 객관성을 도모할 수 있을 것이다.

2. 양질의 데이터 확보

범죄예측의 정확도를 향상시키기 위해서는 양질의 데이터 확보가 가장 중요하다. 이와 마찬가지로 정확한 구속기준을 정립하기 위해서는 구속과 관련한 양질의 데이터가 확보되어야 한다. 그런데 ‘개인정보보호법’ 등에 위반되지 않게 양질의 데이터를 확보하는 것은 쉽지 않다. 그러나 2004년부터 추진된 형사사법 통합 정보체계 구축사업에 의하여 경찰, 검찰, 법원, 법무부 등 형사사법 관련 기관들의 정보를 통합하여 하나의 시스템으로 구축된 형사사법정보시스템(KICS)에는 수사, 기소, 재판, 집행업무를 수행하고 그 결과 생성된 많은 정보와 문서들이 있는데 이들은 빅데이터(Big Data)에 해당한다. 이를 활용하면 구속과 관련한 양질의 데이터를 확보할 수 있을 것이다. 특히 구속영장 결정문에는 많은 양의 정보들을 담고 있다.

3. AI를 활용한 구속 예측 시스템

1) Data Mining을 활용한 구속결정에 관한 패턴 발견

빅데이터의 처리과정은 일반적으로 ① 정보수집 ② 데이터화 ③ 데이터 분석 및 시각화 ④ 새

로운 통찰 또는 가치발견의 순서로 진행된다.[61] 예를 들어 구속영장결정문을 수집하는 것은 정보 수집에 해당되며, 수집된 정보를 일정한 가공과정을 거쳐 데이터로 전환하는 것을 데이터화라고 한다.[62] 수집된 정보가 데이터화를 거쳐 데이터가 되면 그 다음으로 수집·축적된 데이터를 분석하고 시각화하는 과정을 거치게 된다. 이러한 데이터 분석과정을 흔히 '데이터 마이닝'(data mining)이라고 부른다. 데이터 마이닝이란 빅데이터를 분석하여 새로운 통찰이나 가치를 발견하는 과정이라고 말할 수 있다.[63] 데이터 마이닝은 크게 '정형데이터 마이닝'(structured data mining)과 '비정형데이터 마이닝'(unstructured data mining)으로 구분할 수 있다. 정형데이터란 숫자데이터처럼 구조화되어 분석하기 쉬운 데이터를 말한다. 이와 달리 비정형데이터는 그림, 문서, 동영상처럼 그 형태나 구조가 단일하게 구조화되어 있지 않은 데이터를 말한다. 판결이나 구속영장결정은 비정형데이터에 해당한다. 비정형데이터 마이닝에는 '텍스트 마이닝'(text mining), '웹마이닝'(web mining), '오피니언 마이닝'(opinion mining) 등을 언급할 수 있다. 판결나 구속영장결정 등은 텍스트 마이닝으로 분석할 수 있다. 텍스트 마이닝은 대규모의 텍스트 데이터에서 의미 있는 가치나 통찰을 추출하는 것이기 때문이다.[64] 텍스트 마이닝은 자연어처리 기술에 기반하여 비정형 텍스트 데이터에서 유용한 정보를 추출하는 것을 목적으로 하는 기술이다. 다른 정보와의 연계성을 파악하여 방대한 텍스트 뭉치에서 단순한 정보검색 그 이상의 의미 있는 결과를 얻어 낼 수 있다.[65] 수많은 구속영장결정문에서 데이터 마이닝을 사용하여 구속결정에 영향을 주는 의미 있는 패턴(영장기각, 영장발부 등)을 발견하게 될 것이다. 빅데이터를 데이터마이닝하여 획득한 새로운 통찰은 주로 상관성에 기반을 둔 것이다. 어떤 구속요인이 구속결정에 상관성이 있는지를 발견하게 되는 것이다. 이때 '상관성'(correlation)이란 "두 데이터 값 사이의 통계적 관련성을 수량화 한 것"을 말한다.[66] 어떤 구속요인이 구속결정에 얼마나 상관성이 있는지 수치로 나타난다면 구속결정에 개별 구속요인간의 우열이 발생할 것이고 이는 결국 구속기준의 수치화·계량화로 발전할 수 있게 된다.

2) 구속 예측 시스템의 설계 및 활용

보험회사에서는 사고발생률 수준과 사고발생시 피해의 정도를 기준 연령집단과 비교하는 방식으로 보험금 산정을 한다.[67] 이처럼 보험금 산정에 필요한 사고발생률을 계산하는 통계학적 방법과 유사하게 재범의 위험성을 예측하는 방식을 보험 수리적 접근방식(actuarial approach)이라고 한다.[68] 이러한 보험 수리적 접근방식은 기존의 데이터 중에 재범과 관련된 요인에 점수를 부

61) 빅데이터 처리과정에 관한 상세한 설명은 김상락·강만모, "빅데이터 분석 기술의 오늘과 미래", 정보과학지 제32권 1호, 2014.1., 9면 아래 참조.
62) 빅토르 마이어 쇤베르거·케네스 쿠키어, 이지연(옮김), 빅데이터가 만드는 세상, 21세기북스, 2013, 147면.
63) 정용찬, 빅데이터, 케뮤니케이션북스, 2013, 32면 아래 참조.
64) 정용찬, 앞의 글, 42-43면.
65) 한국형사정책연구원, 범죄 빅데이터를 활용한 범죄예방시스템 구축을 위한 예비연구(Ⅰ), 2014, 64면.
66) 빅토르 마이어 쇤베르거·케네스 쿠키어, 이지연(옮김), 빅데이터가 만드는 세상, 21세기북스, 2013, 102면.
67) 이경희, "고연령 자동차보험 계약자의 사고위험 분석", 보험학회지 104권 0호, 한국보험학회, 2015년 10월, 72-79면 참조.
68) Kehl, Danielle et al., Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing, Responsive Communities Initiative, Berkman Klein Center for Internet &

여하고, 설정된 가중치를 적용하여 재범률을 산정한다.[69] 이와 같은 방법으로 특정 범죄자가 가지고 있는 특성 중에 장래에 재범을 저지르는데 기여하는 특정 요인을 입력하여 그 범죄자의 재범률을 산정하는 알고리즘을 재범위험성 예측 알고리즘이라고 한다.[70]

이러한 재범위험성 예측 알고리즘과 같이 보험수리적 접근방식을 이용하여 특정 피의자나 피고인이 가지고 있는 특성 중에 장래에 증거인멸이나 도주우려와 관련 있는 특정 요인을 입력변수로 하여 그 피의자나 피고인의 증거인멸이나 도주할 가능성을 수치로 판단할 수 있는 '구속 예측 알고리즘'을 설계할 수 있을 것이다.

형사사법정보시스템(KICS)에 있는 많은 정보와 문서들 그리고 구속영장결정문 등의 빅데이터를 데이터 마이닝을 통해 분석하게 되면 새로운 패턴이나 상관관계가 발견될 것이다. 이러한 상관관계와 패턴은 구속과 관련된 특정 요인 즉 구속요인들간의 우열을 수량화할 수 있을 것이고 이를 학습한 알고리즘이 구속 예측 알고리즘이 된다.

구속 예측 알고리즘을 이용하여 특정 피의자나 피고인의 특성들을 입력하게 되면 증거인멸이나 도주에 관한 발생률이 산출될 것이다. 산출된 결과물은 구속 여부를 판단하는데 유의미한 예측으로 이어지게 되는데 이를 '구속 예측 시스템'이라고 할 수 있다. 구속 예측 시스템을 활용하게 되면 앞에서 살펴 본 범죄예측 시스템을 활용하게 될 때 발생할 수 있는 문제점들이 발생하게 될 것이다. 즉 구속 예측 시스템을 사용하면 구속과 관련한 개인정보를 침해하는 등 개인의 헌법과 법률상 권리를 침해할 수 있다. 또한 구속 예측 알고리즘도 알고리즘의 불투명성, 편향성, 부정확성 등의 문제가 발생할 것이다. 또한 구속 예측 시스템이 개인과 사회를 통제할 위험성도 있다. 아래에서는 구속 예측 시스템을 활용할 경우 발생할 수 있는 문제점들에 대한 해결방안을 검토하고자 한다.

3) 구속 예측 시스템의 문제점 해결 방안

가. 개인의 privacy와 개인정보보호 방안

구속에 관한 개인정보를 다루게 되면 필연적으로 개인의 권리를 침해할 수 있다. 그렇다고 대상이 되는 개인정보의 수집에 대하여 일일이 개인의 동의를 받거나 법원의 영장을 받는 것은 현실적으로 불가능하다. 이를 개선하기 위해서 2013년 프라이버시에 관한 최대회의인 '제35회 데이터보호 프라이버시 커미셔너 국제회의'에서 채택한 '프로파일링에 대한 결의'를 살펴볼 필요가 있다.[71] 여기서는 프로파일링의 실시조건 6가지가 있는데, ① 프로파일링 실시 이전에 그것에 대한 필요성과 실용성의 점검 등, ② 합법적 목적을 위해 필요한 한도 내에서 행해지는 데이터의 수집과 제한 등, ③ 프로파일링에 의해 야기된 잘못된 결과(긍정적이든 혹은 부정적이든)의 회피 내

Society, Harvard Law School, p.7.

69) Melissa Hamilton, Adventures In Risk: Predicting Violent and Sexual Recidivism in Sentencing Law, 47 Ariz. St. L.J. 1 2015, p.2.

70) 양종모, 앞의 글, 212-213면.

71) 35th International Conference of Data Protection and Privacy Commissioners, Resolution on Profiling, adopted on 24 Sep. 2013.
(출처; https://privacyconference2013.org/Resolutions_and_Declarations).

지 개선을 위한 적당한 조치의 보장 등, ④ 프로파일링이 실시된 것에 대한 가능한 최대한도의 정보제공 등, ⑤ 특히 개인에 대한 중대한 법적 효과를 가지거나 이익이나 지위에 영향을 미치는 그러한 결과에 관한 개인데이터에의 액세스권 및 정정권의 보장 등, ⑥ 모든 프로파일링이 적절한 감독을 받을 것의 보장 등이 열거되었다.[72] 이러한 조건들에 맞도록 범죄예측 시스템에 의한 특정인에 대한 범죄예측도 설계를 할 수 있다. 먼저, 개인정보보호법 제15조 1항 1호에 의해 정보주체의 사전 동의를 반드시 받도록 하여야 할 것이다. 다만, 명백히 정보주체 또는 제3자의 급박한 생명, 신체, 재산의 이익을 위하여 필요하다고 인정되는 경우에는 정보주체의 사전 동의를 받지 않는 예외가 인정될 것이다(동법 제15조 1항 5호).[73] 구속 예측을 위한 특정인에 대한 개인정보수집은 합목적적 범위 내에서 필요최소한에 그쳐야 할 것이고, 잘못된 결과가 발생할 경우(잘못된 긍정이나 잘못된 부정)를 회피 내지 개선할 수 있는 적당한 조치를 사전에 보장하여야 한다. 특히 개인에 대한 중대한 법적 효과를 가지거나 이익이나 지위에 영향을 미치는 그러한 결과에 관한 개인 데이터에의 액세스권 및 정정권의 보장 등을 보장하여야 할 것이다. 모든 개인정보수집은 감독기관의 적절한 감독을 받을 것을 보장하여야 한다.

나. 알고리즘의 문제 해결방안

(1) 외부 감사기관의 설치와 설명가능한 알고리즘

알고리즘의 처리과정을 개발자나 설계자도 정확하게 알지 못하도록 블랙박스처럼 되어 있다는 알고리즘의 불투명성을 해결하는 방법으로는 설명가능한 알고리즘을 개발하는 방법이 있다. 2017년 발표된 아실로마 23원칙[74]이나 유럽연합(EU)의 인공지능윤리 가이드라인 그리고 OECD의 AI 권고안(OECD Recommendation of the Council on Artificial intelligence, 이하 AI원칙)[75]에서도 알고리즘의 설계자나 알고리즘을 사용하는 기업은 그 알고리즘 내부 논리에 대해 감사기관 등에게 설명할 수 있어야 한다는 설명가능성을 명시하고 있다. 이러한 설명가능성을 확보하기 위해서는 사전적·사후적 단계에서 공적인 기관에 의하여 알고리즘에 대한 감사가 필요하다.[76] 따라서 알고리즘의 주체나 피해자를 대신하여 알고리즘의 적절성을 감시하고 심사하는 객

---

72) 차상육, 빅 데이터(Big Data) 환경과 프라이버시 보호, IT와 法 연구 제8집, 2014년(재인용), (출처: https://privacyconference2013.org/web/pageFiles/kcfinder/files/2.%20Profiling%20resolution%20EN%281%29.pdf)

73) 제15조(개인정보의 수집ㆍ이용) ① 개인정보처리자는 다음 각 호의 어느 하나에 해당하는 경우에는 개인정보를 수집할 수 있으며 그 수집 목적의 범위에서 이용할 수 있다.
1. 정보주체의 동의를 받은 경우
2. 법률에 특별한 규정이 있거나 법령상 의무를 준수하기 위하여 불가피한 경우
3. 공공기관이 법령 등에서 정하는 소관 업무의 수행을 위하여 불가피한 경우
4. 정보주체와의 계약의 체결 및 이행을 위하여 불가피하게 필요한 경우
5. 정보주체 또는 그 법정대리인이 의사표시를 할 수 없는 상태에 있거나 주소불명 등으로 사전 동의를 받을 수 없는 경우로서 명백히 정보주체 또는 제3자의 급박한 생명, 신체, 재산의 이익을 위하여 필요하다고 인정되는 경우
6. 개인정보처리자의 정당한 이익을 달성하기 위하여 필요한 경우로서 명백하게 정보주체의 권리보다 우선하는 경우. 이 경우 개인정보처리자의 정당한 이익과 상당한 관련이 있고 합리적인 범위를 초과하지 아니하는 경우에 한한다.

74) 한국인공지능법학회, 인공지능과 법, 박영사, 2020. 66면

75) OECD. Recommendation of the Council on Artificial Intelligence. C/MIN(2019)3/FINAL

관적인 기관이나 조직을 설치하여야 할 것이다.[77]

(2) Daubert 기준

알고리즘의 부정확성 해결방안으로 미국의 과학수사기법의 신뢰성에 관한 Daubert 기준을 사용한 사례가 있다. AI 머신러닝을 사용한 시스템이 오류가 발생하자 그 오류를 수정하는 과정을 반복함으로써 안면인식 시스템의 정확성을 높이고 신뢰성이라는 Daubert 기준을 통과하였다.[78] Daubert 기준처럼 인공지능의 오류를 반복적으로 수정하는 과정을 통해서 알고리즘의 부정확성의 문제를 해결할 수가 있을 것이다. 구속 예측 시스템의 오류를 반복적인 수정과정을 통해서 구속 예측의 정확성을 높여 신병확보와 범죄예방의 효과성이 입증된다면 알고리즘의 부정확성의 문제를 해결할 수가 있을 것이다. 이를 위해서는 오류가 발생하지 않도록 내부적인 통제절차 및 정확성에 대한 인증절차를 마련하는 방안이 필요하다.[79]

(3) 알고리즘의 책무성과 투명성

구속 예측 시스템도 특정인이나 특정 집단을 더 구속하게 되는 등 알고리즘의 편향성의 문제를 해결하기 위해서는 어떤 데이터가 편향된 데이터인지 식별하는 방법과 이러한 편향된 데이터를 학습하여도 편향된 구속 예측을 하지 않는 알고리즘을 개발하는 방법이 필요하다. 하지만 현재로서는 편향된 데이터를 배제하는 방법이나 데이터의 편향성을 극복할 수 있는 알고리즘의 개발은 진행 중이고 다만 알고리즘의 편향성이 얼마나 위험한지를 평가하는 방법이 개발되어 활용되고 있다. 즉 알고리즘에 설계자나 개발자가 의도적으로 또는 무의식적으로 편향성을 알고리즘에 투영하였는지를 알고리즘의 작동방식을 통해서 잠재적인 편향성을 밝히는 방법이다. 이러한 방식은 코넬대학교 연구진과 산업체 연구진이 공동으로 개발하였는데, 먼저 문제된 알고리즘을 모방해서 모델을 만들고 그 모델에 초기 데이터를 투입하여 그 모델의 편향성이 위험한지를 점수로 표시한다. 그리고 나서 초기 데이터를 투입하여 학습한 이후의 모델을 만들어 그 모델의 편향성이 얼마나 위험한지 점수를 내고 이를 이전 모델과 내고 비교한다. 이를 통해서 기초 데이터 중 어떤 것이 점수에 최종 점수에서 영향을 주며 어떤 변수가 중요한지를 추정해 볼 수 있다. 이를 통하여 편향성에 영향을 주는 요인들을 발견하고 그에 대한 조치를 취할 수 있게 된다.[80]

알고리즘의 편향성을 극복하기 위해서는 궁극적으로는 알고리즘의 설계자나 개발자가 알고리즘

76) Michael Veale, Max V. Kleek, and Reuben Binns, Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making, CHI 2018 Paper, Montréal, QC, Canada, 2018. p.1.

77) Tal Z. Zarsky, Transparent Predictions, 2013 U. Ill. L. Rev. 1503, 1515 (2013). at 1558-1562, 1563-1564, 1566.

78) Nawara, John., Machine Learning: Face Recognition Technology Evidence in Criminal Trials, 49 U. Louisville L. Rev. 601(2011). 604, 2011.

79) 황경조/조재웅/서창호, “상호의존정보 기반 공정성을 보장하는 기계학습”, 2019년도 대한전자공학회 추계학술대회 논문집, 2019.11, 667 - 671면.

80) Tan, S., Caruana, R., Hooker, C., & Lou, Y., T., Detecting bias in black-box models using transparent model distilation arXiv preprint arXiv: 1710.0616, 2017.

을 편향되게 설계하지 않도록 하여야 한다(알고리즘 책무성; algorithmic accountability). 이를 위해 설계과정이나 개발과정에서 알고리즘의 요인 및 메트릭스(metrics)를 투명하게 공개하여야 한다(알고리즘의 투명성; algorithmic transparency). 물론 공개가 필요한 알고리즘의 요인 및 메트릭스가 무엇인지는 합의를 통하여 공개할 대상에 대한 표준화가 필요할 것이다. 또한 알고리즘의 투명성이 확보되지 위해서는 신뢰할 수 있는 외부 기관에 의한 감사가 이루어져야 할 것이다.[81)]

다. 구속 예측 시스템의 개인통제기능의 제한

새로운 구속 예측 시스템은 감시 및 통제장치로서 악용될 우려가 있다. 구속 예측 시스템의 이러한 개인통제기능을 제한하기 위해서는 먼저 구속 예측 시스템을 사용하려는 목적과 대상이 사회적 합의에 의하여 구체화되어야 한다. 구속 예측 시스템을 사용하려는 목적이 범죄예방을 통한 피해자 보호, 피의자 또는 피고인의 인권 보호, 공권력의 효율적인 집행 등을 고려하여 구체적으로 설정되어야 한다. 범죄예방, 피해자의 보호, 피고인이나 피의자의 권리 보장, 국가기능의 효율성 등의 법익이 상호 충돌되지 않도록 비례성의 원칙을 적용하여 하고 관련된 자의 권리침해를 최소화하도록 구속 예측 시스템을 사용하여야 할 것이다(최소침해의 원칙). 구속 예측 시스템을 대상자의 도주나 증거인멸의 재범위험성에 따라 개별적으로 사용하여야 한다(개별성의 원칙). 그리고 구속 예측 시스템이 국민에 대한 감시 및 통제 시스템으로 작동하지 않는지를 시민단체나 공적 감시기관을 통해서 정기적으로 감사되어야 할 것이다.

4. 구속영장에서 개별 구속요인들의 평가방식

특정 피의자나 피고인에 대하여 구속 예측 시스템을 사용하여 산출한 결과물을 가지고 바로 구속을 집행한다면 국가기능의 효율성에서는 높은 점수를 받을 수 있겠지만 구속 예측 시스템의 대상이 된 피의자 또는 피고인의 입장에서는 그러한 결과가 도출되었는지를 알아야 구속결정에 대한 수긍을 할 수가 있을 것이다. 또한 일반 국민의 입장에서도 구속 결정이 공정한 기준에 의하여 집행되었다는 것을 확인할 수 있을 때 공권력 행사에 대한 정당성을 인정할 수가 있을 것이다. 즉 구속 결정에 대한 대상자와 일반 국민의 수긍과 인정이 있어야 하는데 이것은 구속영장에 구속 예측 시스템의 결과가 구속결정으로 이어졌는지를 가시적으로 표시되어야 한다. 이를 위해서는 구속 예측 시스템의 알고리즘이 설명가능한 알고리즘으로 전제되어야 하고 구속 예측 시스템으로 산출한 증거인멸이나 도주의 위험성이 수치화로 표현되고 이들에 어떠한 구속요인들이 작동했는지를 시각화하는 것이 필요하다. 따라서 아래에서는 구속영장에 구속결정에 고려한 구속요인이 구속 예측 시스템의 결과와 연결되는지를 구속영장에 설시하는 방식에 관하여 검토하고자 한다.

---

81) 주현경/정채연, “범죄예측 및 형사사법절차에서 알고리즘 편향성 문제와 인공지능의 활용을 위한 규범 설계”, 법학논총 제27권 제1호, 2020, 134면; Nicholas Diakopoulos, Algorithmic Accountability Reporting: On the Investigation of Black Boxes, Two Center for Digital Journalism - A Tow/Knight Brief, Columbia Journalism School, 2013. p.28.

1) 단순나열 방식

구속사유를 판단할 때에 고려한 구속요인을 구속영장에 적시하여 어떤 근거로 구속영장이 발부되거나 기각되었는지를 추론할 수 있는 방식이다. 단순나열방식은 구속사유가 존재하는지 여부를 판단할 때 구체적으로 어떠한 구속요인이 참작이 되었는지를 구속영장에 나타내는 것만으로도 과거 아무런 구속요인에 대한 판단 없이 일률적으로 '증거인멸 또는 도망의 위험'이라는 획일적인 구속사유를 기재하는 방식보다는 진보된 방식임에는 틀림이 없다. 즉 앞에서 언급한 것처럼 법관이 구속여부를 판단할 때 구속요인들을 실제로 고려하였는지 여부를 외부에서는 전혀 알 수가 없었다는 문제가 있었는데, 단순나열방식에서는 구속여부를 판단하는데 어떤 구속요인들이 고려되었는지를 적시할 수 있기 때문에 이전보다는 진일보되었다는 것이다.

그러나 이러한 방식은 구체적 사건에서 구속여부를 판단하는데 참고자료이상의 의미를 부여하지 못하고 실질적인 준거를 제시할 수 없다는 점에서 문제가 있다. 구속요인을 단순히 나열하여 고려한 요인들을 체크만 하고 이에 대한 평가는 전적으로 법관에게 맡기는 것은 기존의 구속관행과 별다른 것이 없기 때문이다. 구속에서 중요한 것은 구속요인을 단순히 나열하고 적시하는 것이 아니라 구속요인을 구체적으로 어떻게 평가하고 구속요인 사이의 경중이나 우선순위를 판단하고 이것이 어떻게 작용하여 구속여부를 결정하는가에 있다. 그러므로 구속요인의 단순나열방식은 구속여부를 결정하는데 있어서 법관이 어떤 구속요인을 얼마만큼 고려하였는지를 나타낼 수 없기 때문에 구속여부에 대한 명확한 근거를 제시하는 것에는 한계가 있을 수밖에 없다.

【단순나열 방식-증거인멸의 염려】

번호	증거인멸의 염려를 판단하는 요인들	O,X
1	사안이 중요한가(사건의 중대성)?	
2	증거인멸의 대상이 되는 결정적 증거가 존재하는가?	
3	피의자의 성행, 지능과 환경으로 보아 증거인멸의 가능성이 있는가?	
4	증거를 인멸할 공범이 존재하는가?	
5	피해자 및 중요 참고인 등에 대한 위해의 우려가 있는가?	
6	피의자측이 피해자 등 증인에 대하여 어느 정도의 압력이나 영향력을 행사할 수 있는가?	
7	수사 협조 등 범행 후의 정황으로 보아 증거인멸의 가능성 있는가?	
8	범죄의 전력을 통해 증거인멸의 가능성이 있는가?	
종합	구속사유를 판단하는데 고려한 구속요인들의 번호:	

【단순나열 방식-도망의 위험】

번호	도망의 위험을 판단하는 요인들	O,X
1	사안이 중요한가(사건의 중대성)?	
2	재범의 위험성이 있는가(재범의 위험성)?	
3	범행의 동기, 수단과 결과로 보아 도망의 가능성이 있는가?	
4	피의자의 직업, 재산, 교우, 조직·지역 사회정착성, 사회적 환경으로 보아 도망의 가능성이 있는가?	
5	주거가 부정한가?	
6	국외 근거지의 존재 여부, 출국 행태로 보아 도망의 가능성이 있는가?	
7	수사 협조 등 범행 후의 정황으로 보아 도망의 위험성이 있는가?	
8	범죄 전력으로 보아 도망의 가능성이 있는가?	
9	자수를 했는가?	
10	피해자와의 관계, 피해 회복 및 합의를 했는가?	
종합	구속사유를 판단하는데 고려한 구속요인들의 번호:	

2) 개별 구속요인에 점수 부여방식

개별 구속요인들에게 점수를 부여하여 일정 점수 이상이면 구속여부를 결정하는 준거로 사용하는 방법이 있는데, 이 방법은 구속요인을 단순히 나열하는 것보다 구속여부를 결정하는데 구체적으로 어떠한 구속요인들이 고려되었다는 것을 보여줄 수 있으며 또한 일정 점수 이상의 경우에 구속여부를 결정하였기 때문에 구속결정에 정당성을 부여할 수도 있다. 여기에는 모든 구속요인들에게 동일한 점수를 부여하는 방법과 개별 구속요인들간의 우열을 인정하여 차등 점수를 부여하는 방법이 있다.

(1) 개별 구속요인들에게 동일한 점수 부여방식

먼저, 개별 구속요인들에게 동일한 점수를 부여하는 방안은 구체적 사건에서 개별 구속사유의 구속요인의 점수의 합계가 일정점수 이상이면 필요적 구속으로 판단하는 방법이다. 이를 위해서는 개별 구속요인들에게 동일하게 점수를 부여하는 것이 타당한지를 검토하여야 하고, 또한 개별 구속요인의 점수의 합계가 어느 정도의 점수가 되면 필요적 구속으로 간주할 것인지를 결정하여야 할 것이다.

【동일점수부여방식-증거인멸의 염려】

번호	증거인멸의 염려를 판단하는 요인들	O(1점)	X(0점)
1	사안이 중요한가(사건의 중대성)?		
2	증거인멸의 대상이 되는 결정적 증거가 존재하는가?		
3	증거를 인멸할 공범이 존재하는가?		
4	피의자의 성행, 지능과 환경으로 보아 증거인멸의 가능성이 있는가?		
5	피해자 및 중요 참고인 등에 대한 위해의 우려가 있는가?		
6	피의자측이 피해자 등 증인에 대하여 어느 정도의 압력이나 영향력을 행사할 수 있는가?		
7	수사 협조 등 범행 후의 정황으로 보아 증거인멸의 가능성 있는가?		
8	범죄의 전력을 통해 증거인멸의 가능성이 있는가?		
총점	1~3점: 증거인멸의 가능성 미약- 필요적 불구속 4~6점: 증거인멸의 가능성 있음- 구속여부 재량판단 가능함. 7점 이상: 증거인멸가능성 높음- 필요적 구속		

【동일점수부여방식-도망의 위험】

번호	도망의 위험을 판단하는 요인들	O(1점)	X(0점)
1	사안이 중요한가(사건의 중대성)?		
2	재범의 위험성이 있는가(재범의 위험성)?		
3	범행의 동기, 수단과 결과로 보아 도망의 가능성이 있는가?		
4	피의자의 직업, 재산, 교우, 조직·지역 사회정착성, 사회적 환경으로 보아 도망의 가능성이 있는가?		
5	주거가 부정한가(주거의 부정성)?		
6	국외 근거지의 존재 여부, 출국 행태로 보아 도망의 가능성이 있는가?		
7	수사 협조 등 범행 후의 정황으로 보아 도망의 위험성이 있는가?		
8	범죄 전력으로 보아 도망의 가능성이 있는가?		
9	자수를 했는가?	-1점	1점
10	피해자와의 관계, 피해 회복 및 합의를 했는가?	-1점	1점
총점	1~3점: 도망의 가능성 미약- 필요적 불구속 4~6점: 도망의 위험성 있음- 구속여부 재량판단 가능함. 7점 이상: 도망의 위험성 높음- 필요적 구속		

개별 구속요인의 총점이 1~3점이면 도망의 가능성이 미약한 것으로 필요적 불구속이다. 그러나 총점이 4~6점인 경우는 도망의 위험성이 있으나 법관의 재량 판단이 가능한 부분이어서 구속여부 결정은 법관의 법적 양심에 맡겨두는 것이 타당할 것이다. 그러나 총점이 7점 이상인 경우에는 도망의 위험성이 높아 필요적으로 구속을 하여야 할 경우이다.

(2) 개별 구속요인들에게 차등점수 부여방식

개별 구속요인들 중에 구속사유를 결정하는데 중요도에 따라 가중치를 두어 차등적인 점수를 부여하여 방법이다. 이 경우도 고려된 구속요인들 점수의 합이 일정 점수 이상이면 필요적 구속으로 결정하는 방법이다. 이것은 개별 범죄의 구속여부를 결정하는데 개별 구속요인들이 상대적으로 중요도가 다르다는 것을 의미한다. 즉, 증거인멸의 염려와 관련성이 있을 것으로 보이는 자

백여부는 일부범죄에 국한되어 구속영장발부여부를 결정하는 요인으로 확인되었고, 전반적으로 구속영장 기각에 영향을 주는 요인으로는 동종전과 또는 동종전과 중 집행유예 이상의 여부와 사전영장여부로 나타난다. 이 결과는 오히려 재범의 위험성이 구속영장발부여부를 결정짓는데 더 중요한 고려요인임을 의미한다.[82)]

이러한 방법을 채택하기 위하여서는 각 개별 구속요인들에게 점수를 어느 정도 부여할 것인지를 결정하여야 할 것이다. 그러기 위해서는 구속요인들간에 우열을 인정하고 어떤 기준에서 가중치를 두어야 할지가 결정되어야 한다.

【차등점수부여방식- 증거인멸의 염려】

번호	증거인멸의 염려를 판단하는 요인들	상(2점)	중(1점)	하(0점)
1	사안이 중요한가(사건의 중대성)?			
2	증거인멸의 대상이 되는 결정적 증거가 존재하는가?			
3	증거를 인멸할 공범이 존재하는가?			
4	피의자의 성행, 지능과 환경으로 보아 증거인멸의 가능성이 있는가?			
5	피해자 및 중요 참고인 등에 대한 위해의 우려가 있는가?			
6	피의자측이 피해자 등 증인에 대하여 어느 정도의 압력이나 영향력을 행사할 수 있는가?			
7	수사 협조 등 범행 후의 정황으로 보아 증거인멸의 가능성 있는가?			
8	범죄의 전력을 통해 증거인멸의 가능성이 있는가?			
총점	1~3점: 증거인멸의 가능성 미약- 필요적 불구속 4~7 : 증거인멸의 가능성 있음- 구속여부 재량판단 가능함. 8 이상: 증거인멸가능성 높음- 필요적 구속			

【차등점수부여방식- 도망의 위험】

번호	도망의 위험을 판단하는 요인들	상(2점)	중(1점)	하(0점)
1	사안이 중요한가(사건의 중대성)?			
2	재범의 위험성이 있는가(재범의 위험성)?			
3	범행의 동기, 수단과 결과로 보아 도망의 가능성이 있는가?			
4	피의자의 직업, 재산, 교우, 조직·지역 사회정착성, 사회적 환경으로 보아 도망의 가능성이 있는가?			
5	주거가 부정한가?			
6	국외 근거지의 존재 여부, 출국 행태로 보아 도망의 가능성이 있는가?			
7	수사 협조 등 범행 후의 정황으로 보아 도망의 위험성이 있는가?	수사에 협조함(-1점)/안함(1점)		
8	범죄 전력으로 보아 도망의 가능성이 있는가?			
9	자수를 했는가?	자수 안했다(1점)/했다(-2점)		
10	피해자와의 관계, 피해 회복 및 합의를 했는가?	합의 안했다(1점)/했다(-2점)		
총점	1~4점: 도망의 가능성 미약- 필요적 불구속 5~9점: 도망의 위험성 있음- 구속여부 재량판단 가능함. 10점 이상: 도망의 위험성 높음- 필요적 구속			

위의 경우처럼 개별 구속요인에 동일한 점수를 부여하고 각 경우에 '있다'와 '없다'로 나누어 각

82) 김정훈.이나래.이계민, 앞의 글, 1233면.

각 1점과 0점을 부여하는 방법도 있지만, 개별사안마다 상, 중, 하로 나누어 각각 2점, 1점, 0점의 방법도 있을 수 있다. 그러나 이 경우에 각 개별요인들의 점수를 동일하게 산정하는 것이 타당한지 여부가 논의될 수 있을 것이다. 각 개별요인들의 점수 산정을 '있다'와 '없다'로 2단계 또는 상, 중, 하로 3단계로 일률적으로 나누는 것이 적정한지도 검토되어야 할 것이다.

6. 기대효과와 활용방안

1) 기대효과

가. 국민의 법감정에 부합 및 사법불신의 해소

구속요인들 간의 우열을 판단하고 이들을 수치화·계량화함으로써 실제 사건에서 구속여부를 검찰이나 경찰과 같은 수사기관은 물론이고 일반국민들까지 예측할 수 있게 된다면, 국민들이 인신구속이 될지도 모른다는 두려움에서 벗어나기 위해 가장 최근에 개업한 '따끈따끈한' 전관출신의 변호사를 찾아 고액을 지불하지 않아도 될 것이다. 이를 통해서 전관예우의 관행을 줄임은 물론, 국민들이 법률생활에서 예측가능성과 법적 안정성을 높여 법을 몰라서 발생하는 억울한 피해를 줄일 수가 있을 것이다. 그리고 재벌총수나 정치인이 온갖 비리와 횡령, 뇌물 등 범죄사건에 연루되고도 구속되지 않거나 그 재판의 결과로 집행유예 등 솜방망이 처벌을 받는 것을 보아 온 상당수의 국민들은 재벌총수나 정치인 등에 대한 구속기준과 처벌기준이 일반국민들에 대한 기준들과 차별적으로 적용되는 것처럼 느껴왔던 것이 사실이다. 그러나 객관적 구속기준이 정립된다면 국민이 이해하고 공감할 수 있는 사법을 실현할 수 있을 것이고 '유전불구속 무전구속'이나 '유전무죄 무전유죄'로 대표되는 사법에 대한 불신도 감소하게 될 것이다.

나. 통일적 구속기준의 정립을 통한 검찰과 법원의 업무의 효율성 증진 및 갈등해소

구속사유는 검사나 법관의 주관적 추측이나 염려가 아니라 구체적 사실을 기초로 객관화되어야 한다. 따라서 통상인의 합리적 판단에 의하여 수긍할 수 있을 정도의 것임을 요한다. 여기에서 구속영장의 발부기준의 명확화와 함께 예측가능성을 담보하기 위해선 검찰과 법원의 각기 다른 영장기준들을 통일화하는 것이 중요하다. 각 기관마다 각기 다른 영장기준을 적용하고 있는 것은 국민들에게 혼란을 가중하는 것이다. 법원의 경우에는 검찰의 구체적 기준을 참고하여야 할 것이고 검찰도 법원의 구체적 기준을 참작하여야 할 것이다. 궁극적으로는 이들 기관들이 동일한 기준을 적용하여 국민들에게 혼란을 주어서는 안 될 것이다.[83] 그리고 동일한 기준을 각 기관의 개별적인 법규에 규정하는 것보다 형사소송법에 규정하는 것이 타당할 것이다. 그러나 법률에 규정하는 것도 한계가 있기 때문에 형사소송법규칙에 규정하는 것이 타당할 것이다.

검찰과 법원이 개별기준을 가지고 구속제도를 운영하는 것은 불필요한 사법행정적 업무를 발생

83) 김병수, 앞의 글, 192면.

시켜 제도의 효율적 운용을 저해하는 요인이 되어 왔다. 지금처럼 구속영장의 기각률이 높다는 것은 법원측에서는 불필요한 구속영장을 검찰이 신청한 것이 된다. 이는 법원과 검찰간에 구속기준이 공유되었다면 검찰은 구속영장이 기각될 것이 예상되는 사안에서는 신청을 하지 않음으로써 불필요한 인력과 시간의 낭비를 막을 수 있다는 것을 의미한다. 그러므로 법원과 검찰 그리고 국민의 합의에 의해 계량화된 구속기준이 확립될 경우 수사기관에서도 이 기준에 따라 구속영장을 신청할 것이기 때문에 구속영장의 발부와 관련하여 발생할 수 있는 인력의 낭비와 불필요한 신경전 등의 문제가 크게 줄어들 수 있다. 이를 통하여 불필요한 구속영장의 신청 및 청구에 따른 사법행정업무를 줄이고, 피의자에 대한 실질적인 심문이 가능하게 되며, 피의자의 인권보호에도 기여하게 될 것이다. 또한 이 기준에 따라 구속이 되지 않을 것으로 예상되는 사람에 대한 구속영장의 신청을 하지 않을 가능성이 높아지면서 자연히 영장의 청구 자체가 신중해질 것으로 예상된다.[84] 그렇게 된다면 현재와 같은 검찰과 법원간의 구속영장청구와 기각을 둘러싼 갈등을 해소하는데 기여할 수 있다.

다. 시민법교육에 활용

과거 어려운 법적 용어와 논리에 의해서만 이루어진 법원의 판단결과에 대해 얼마나 공정성을 갖는지 의문을 제기하는 경우가 많았었고 특히 구속영장 판단이 국민의 일반적인 법감정과 거리가 있을 경우에는 이러한 공정성의 문제가 더욱 부각되었다. 법치주의의 근간이 국민들의 법에 대한 신뢰에서 비롯된다는 점을 고려해보았을 때 사법제도에 대한 신뢰 저하는 심각한 문제라고 볼 수 있다. 법에 대한 신뢰가 낮아지는 것은 복합적인 원인을 통해 나타나는 현상이다. 법자체가 부당하거나 법집행이 불공정하게 이루어지는 현상이 반복되면서 불신이 쌓이는 경우도 있지만 제도와 제도운영에 대한 오해나 편견에서 비롯된 측면이 적지 않다는 측면에서도 국민이 예측가능하도록 구속기준을 계량화한다면 사법에 대한 불신을 불식시키는 효과뿐만 아니라 시민을 대상으로 한 법교육의 효과를 가져 올 수 있다. 계량화된 구속기준을 통해 시민들이 스스로 구속여부를 판단하는 것이 가능하여지게 되면 사법제도에 대하여 보다 더 친숙한 태도를 갖게 되고, 사법에 대한 이해심도 키우게 된다. 구속여부에 대하여 예측할 수 있게 된 시민들은 법원의 영장결정에 대하여 승복하기가 더 쉬울 것이고 이는 자연스럽게 국민적 통합의 계기가 될 것이다. 계량화된 구속기준은 사법을 시민들에게 가까이 다가가게 만들고 이해시키는 시민일반에 대한 교육적 효과도 가지게 된다.

2) 활용방안

가. 구속과 양형판단의 준거기준을 제시

구속요인들의 수치화·계량화는 개별사건에서 구속여부를 예측할 수 있는 자료로서도 사용되지만

---

84) 신의기, “검찰과 법원의 구속영장신청 및 발부기준 차이의 해결방안”, 형사법의 신동향 제26호, 2010. 6., 138-139면.

양형판단과 관련하여 양형요인들간의 우열 여부를 판단할 수 있는 준거를 제공하게 되어 양형자료조사에 구체적인 해석기준을 제공함을 물론 현재 진행 중인 양형기준제도를 보완할 수 있는 획기적인 전환의 계기가 될 것이다. 현재 진행 중인 양형기준에 관한 논의들 중에는 현행 양형기준에 대하여 양형기준의 권고적 효력의 문제, 지나치게 넓은 형량범위의 문제, 양형인자의 평가방법의 문제, 집행유예의 기준에 관한 문제 등으로 많은 비판들이 있다.[85] 이 중에서 특히 양형인자의 동등취급과 관련한 문제를 해결하는데 구속요인에 대한 계량화는 실질적인 도움을 줄 수 있을 것이다. 양형기준에서도 기존의 양형인자들을 단순히 나열만 한 채 이에 대한 평가는 전적으로 법관에게 맡기고 있어[86] 개선이 시급한데, 구속요인의 계량화는 양형기준에서 양형인자를 구체적으로 어떻게 평가하고 양형요인 사이의 경중이나 우선순위를 정하는데 실질적인 도움을 줄 것이다. 이러한 구속요인과 양형요인의 수치화·계량화가 성공을 한다면, 이것을 외국에도 소개하여 각 나라의 구속과 양형판단에 대한 준거기준을 제시할 수 있을 것으로 기대된다.

나. 범죄예방 및 범죄대응정책의 근거자료로 활용

구속과 관련된 형사사법정보를 인공지능에 기반한 범죄예측시스템으로 분석하게 되면 구속인자 이외에도 개별범죄의 경향성 및 개별범죄에 영향을 미치는 범죄관련인자들도 발견하게 될 것이다. 이들은 범죄예방과 범죄대응전략의 근거자료로 활용할 수 있을 것이다.

다. 예방적 개입의 판단기준으로 활용

정신질환자에 의한 진주방화사건 이후로 범죄에 대하여 예방적 개입이 필요함을 절실히 실감하고 있다. 정신장애범죄 뿐만 아니라 아동학대, 노인학대, 학교폭력 등에서도 예방적 개입이 필요하다. 하지만 언제 어떻게 개입하여야 할지 판단기준이 제대로 마련되어 있지 않다. 그러나 계량화 수치화된 구속인자와 같이 이들 예방적 개입이 필요한 부분에도 관련인자들의 계량화 수치화가 예방적 개입의 판단기준을 정립하는데 도움을 줄 것이다.

Ⅴ. 나가며

'증거인멸', '도주우려' 등 현행 구속사유(형사소송법 제70조)가 너무 추상적이어서 구체적 사건에

85) 최호진, "특집 : 새로운 범죄현상과 형사법의 현대적 과제: 양형기준의 합리성 검토와 개선방향", 비교형사법연구 제11권 제2호, 한국비교형사법학회, 2009, 88-90면; 김현철, 살인범죄 양형기준안 지정토론문, 양형위원회, 2008.11, 262-265면; 신현호, 뇌물범죄 양형기준안 지정토론문, 양형위원회, 2008.11, 297면; 이광수, 횡령·배임범죄 양형기준안 지정토론문, 양형위원회, 2009.2, 217-220면; 한상희, 뇌물범죄 양형기준안 지정토론문, 양형위원회, 2008.11, 290면; 이현정·임웅, "양형위원회의 양형기준안에 대한 비판적 고찰", 성균관법학 제21권 제1호, 2009.4, 372-376면; 오영근, "바람직한 양형기준의 방식", 형사정책연구 제20권 제1호, 한국형사정책연구원, 2009, 362-363면; 김영종, 위증·무고범죄 양형기준안 지정토론문, 양형기준안 제2차 공청회자료집, 양형위원회, 2009.2, 264면; 이주형, "최소 양형기준에 대한 비판적 고찰", 형사법의 신동향 통권 제21호, 2009.8, 73면.

86) 오영근, 앞의 글, 362-363면.

서 정확한 구속기준이 뭔지 알기도 어렵고 예측하기도 어렵기 때문이다. 게다가 법률전문가인 검찰과 법원도 국민들이 공감할 수 있는 명확한 판단기준을 제시하고 못하고 있다. 이 때문에 국민들은 어떤 경우에 구속이 되는지를 몰라 인신구속의 두려움을 해소하기 위하여 비싼 수임료를 부담해서라도 전관출신의 변호사를 찾고 있는 것이 현실이며, 재벌총수와 정치인 등 돈 있는 사람과 힘없고 돈 없는 서민의 구속기준이 다르다고 생각하여 '유전불구속 무전구속'이란 말까지 하여 왔다. 국민들의 법률생활을 불안정하게 하고 예측가능성을 곤란하게 하며 형평성에 의구심까지 갖게 하는 구속기준을 재정립하는 것은 시대의 요청인 것이다.

과거에는 구속판단은 사실상 법관의 감(感)에 의한 주관적 판단에 의존하였던 것이 사실이다. 그러나 앞으로의 구속판단은 객관적인 데이터에 근거한 객관적인 판단이어야 할 것이다. 그러기 위해서는 수사자료, 조서, 영장, 공소장, 판결문 등의 같이 형사사법기관들이 형사사법업무와 관련하여 축적한 형사사법정보를 바탕으로 객관적인 판단을 하여야 할 것이다. 이러한 정보들은 방대하여 빅데이터에 속한다. 오늘날 이러한 빅데이터는 인간이 수집, 분석 및 처리하는데 한계가 있으므로 머신러닝 및 딥러닝 기술에 근거한 인공지능이 수집, 분석 및 처리를 하고 있다. 이처럼 방대한 사법정보를 인공지능이 처리하여 구속기준을 재정립할 수 있다면 법관의 감에 의한 주관적인 구속판단에서 데이터에 기반한 객관적인 구속판단으로 구속판단의 객관성을 도모할 수 있을 것이다. 우선 정확한 구속기준을 정립하기 위해서는 구속과 관련한 양질의 데이터가 확보되어야 한다. 형사사법정보시스템(KICS)에는 구속영장 결정문 등과 같은 양질의 빅데이터를 활용할 수가 있을 것이다. 이러한 빅데이터를 데이터 마이닝을 사용하여 분석하면 구속결정에 영향을 주는 의미 있는 패턴(영장기각, 영장발부 등)을 발견하게 될 것이다. 이를 통하여 어떤 구속요인이 구속결정에 얼마나 상관성이 있는지 수치로 나타난다면 구속결정에 개별 구속요인간의 우열이 발생할 것이고 이는 결국 구속기준의 수치화·계량화로 발전할 수 있게 될 것이다. 그리고 구속사유를 판단할 때에 고려한 구속요인을 구속영장에 적시하여 어떤 근거로 구속영장이 발부되거나 기각되었는지를 추론할 수 있는 방식이 필요한데, 개별 구속요인의 단순 나열방식과 점수부여방식이 있다. 개별 구속요인의 단순나열방식은 현행 구속사유 판단방식과 큰 차이가 없으므로 점수부여방식을 선택하였는데 여기에도 개별구속요인에 동일점수 부여방식과 차등점수부여방식이 있다. 차등점수부여방식이 개별구속요인의 판단에 적합할 것으로 사료된다. 그러나 개별 구속요인간의 우열화와 계량화에 관한 구체적인 논의는 국민과 검찰, 법원 그리고 학계가 머리를 맞대고 협력을 하여야 할 것이다.

[참고문헌]

구본진, “현행 피의자 구속제도의 위헌적 요인”, 저스티스 제125호, 2011.

강동범, “구속영장재판에 대한 불복 가부”, 형사법연구 제21권 3호, 2009.

김경수, “영장항고제도 도입 필요성에 관한 연구”, 형사정책연구 제84호, 2010.

김병수, “구속요인의 계량화를 통한 구속기준 정립”, 형사법연구 제24권 2호, 2012,

김성룡, “삼성전자 이재용 부회장 구속영장 발부를 둘러싼 법적 쟁점”, 형사법연구 제29권 제2호, 2017.

김태완, “데이터 마이닝 기법을 활용한 청소년 재범 관련요인 연구”, 교정복지연구(제35호), 한국교정복지학회(2014. 12).

김현수, “피의자 구속의 합리적 기준에 관한 제언”, 인권과 정의 Vol.390.

김현철, 살인범죄 양형기준안 지정토론문, 양형위원회, 2008.11.

노정환, "한、중 피의자 인신구속제도에 관한 연구”, 법학논총 제17권, 2007.

박남미, “구속영장 청구 및 발부에 있어서 구속사유 판단 기준에 대한 고찰”, 형사법의 신동향 59권, 2018.

변종필, “구속제도의 이론과 실제”, 형사정책연구 제82호, 2010.

신의기, “검찰과 법원의 구속영장신청 및 발부기준 차이의 해결방안”, 형사법의 신동향 제26호, 2010.

신현호, 뇌물범죄 양형기준안 지정토론문, 양형위원회, 2008.11.

이동희, “한국과 일본의 체포 구속제도에 대한 비교연구 -체포 구속의 법제 및 실무운용을 중심으로-”, 비교형사법연구 제11권 1호, 2009.

양종모, “인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안”, 법조 제723권, 법조협회, 2017.

양종모, “형사사법절차 전자화와 빅 데이터를 이용한 양형합리화 방안 모색”, 홍익법학 제17권 1호, 2016.2.

양종모, “인공지능 이용 범죄예측 기법과 불심검문 등에의 적용에 관한 고찰”, 형사법의 신동향 제51권, 2016.6.

윤상현/신상화, “범죄 예측의 현재와 미래- 한국형 범죄예측 시스템 개선.도입에 대한 제언 -”, 한국공안행정확회보, 제71호, 2018.

윤상연, “범죄예측 기술 어디까지 왔나?,” 치안정책리뷰, 제60호, 2018. 6.

윤해성 등, 범죄 빅데이터를 활용한 범죄예방시스템 구축을 위한 예비 연구(I). 형사정책연구원, 2014.

임승옥 외, “범죄 발생환경, 행동패턴 및 심리정보 등 융합정보 적용형 엔트로피 필터링 예측분석기반의 실시간 범죄 예측· 예방 시스템 개발,” 정보통신· 방송 연구 개발사업 최종보고서, 전자부품연구원, 2016. 10.

이광수, 횡령·배임범죄 양형기준안 지정토론문, 양형위원회, 2009.2

이수정, “성인 판결전 조사 활용, 재범 위험성 평가도구 개발”, 동향/연구보고서, 법무부(2010).

이승주, “현행 구속사유 해석을 둘러싼 혼란과 그 개선방안”, 저스티스 제167권, 2018.
이승준, “영장발부에 있어 구속사유 판단의 문제-개정 형사소송법 제70조 제2항을 중심으로-”, 형사정책연구 제19권 제4호, 2008.
이완규, “독일의 구속영장 심사와 우리나라 실무에의 시사점”, 형사법의 신동향 통권 제7호, 2007.
이태억, “[이태억의 과학에세이] 스몰데이터 AI와 뉴턴의 사과”, IT조선, 2018. 7. 7.
조국, “영미 체포, 구속제도의 특징과 그 비교법적 함의”, 비교형사법연구 제11권 1호, 2009.
치안정책연구소. 범죄예측 기술 어디까지 왔나?, 치안정책리뷰, 제 60호, 2018.
최호진, “특집 : 새로운 범죄현상과 형사법의 현대적 과제: 양형기준의 합리성 검토와 개선방향”, 비교형사법연구 제11권 제2호, 한국비교형사법학회, 2009.
크리스토퍼 스타이너/박지유(역),『알고리즘으로 세상을 지배하라』, 에이콘(2016).
한국연구재단, 2013년 교육부 우수논문지원사업성과사례집, 2013.
한국전자통신연구원, “위험 상황 초기 인지를 위한 ICT 기반의 범죄 위험도 예측 및 대응 기술 개발,” 한국전자통신연구원 사업계획서, 2018.
한국전자통신연구원, “지능형 치안서비스 기술 동향”, 전자통신동향분석 제34권 제1호, 2019. 2.
황경조/조재웅/서창호, “상호의존정보 기반 공정성을 보장하는 기계학습”, 2019년도 대한전자공학회 추계학술대회 논문집, 2019.11.

Andrea Roth, Trial by Machine, 104 Geo. L. J. 1245 (2016).
C. Francescani, “NYPD Expands Surveillance net to Fight Crime as Well as Terrorism,” Reuters, June 21, 2013.
Harthorne M., City : A Few streets Responsible for Half of seatle's Crime; komonews.com/archive/city-a-few-streets-responsible-for-half-of-seatles-crime .
Holmes, S. T., Tewksbury, R., & Holmes, R. M. (1999). Fractured Identity Syndrome. Journal of Contemporary Criminal Justice, 13(3).
Jeffrey S. Paul and Thomas M. Joiner, Integration of Centralized Intelligence with Geographic Information Systems, Geography and Public Safety, Oct. 2011.
Jennifer Bachner, Predictive Policing: Predictive Policing: Preventing Crime with Data and Analytics, Washington, DC: IBM Center for The Business of Government, 2013.
Nath, S.V., Crime Pattern Detection Using Data Mining, Oracle Corporation, 2006.
Perry, W. L. (2013), Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations, RAND Corporation.
Singer, S. D., & Hensley, C. (2004). Applying Social Learning Theory to Childhood and Adolescent Firesetting: Can It Lead to Serial Murder?. International journal of

offender therapy and comparative criminology: official organ of the Association for Psychiatric Treatment of Offenders (APTO), 48(4).

Sherman, L.W., Gartin, P.R., & Buerger, M.E., "Hot Spots of Predatory Crime: Routine Activities and the Criminology of Place," Criminology, 27, 1989.

Tal Z. Zarsky, Transparent Predictions, 2013 U. Ill. L. Rev. 1503, 1515 (2013).

Whitman, T. A., & Akutagawa, D. (2004). Riddles in serial murder: A synthesis. Aggression and Violent Behavior, 9.

Wright, J., & Hensley, C. (2003). From Animal Cruelty to Serial Murder: Applying the Graduation Hypothesis. International Journal of Offender Therapy and Comparative Criminology: official organ of the Association for Psychiatric Treatment of Offenders(APTO), 47(1).

Zach Friend, Predictive Policing: Using Technology to Reduce Crime, FBI Law Enforcement Bulletin(Feb. 8, 2018).

< 토 론 문 >

“AI를 활용한 구속기준의 재정립”에 대한 토론문

토론자: 전남대학교 법학전문대학원 최병천 교수

AI를 활용한 구속기준의 재정립이라는 발표문을 통해 낯설게 당면하고 있음에도 불구하고, 우리가 반드시 도전해야 할 것으로 생각되는 주제에 접근할 수 있는 기회를 주신 발표자에게 감사의 말씀을 드립니다.

미지의 세계인 AI를 활용한 구속기준의 재정립을 모색하기 위해서 유사 분야라고 할 수 있는 AI를 이용한 범죄예측 시스템의 현황과 그 문제점을 개관해 본 것은 AI를 활용한 구속기준에 관한 연구의 방향을 가늠하고 그 내용을 구체화할 수 있는 바람직하고 합리적인 연구방법이라고 생각됩니다.

AI의 세계를 전혀 알지 못하는 사람으로서 발표문에 대해 언급하게 되어 토론이 부정확하고 오류가 많을 것으로 생각됩니다.

구속결정에 관한 정보수집 및 데이터화에 관하여, 발표자는 구속영장결정문에서 데이터마이닝을 사용하여 구속결정에 영향을 주는 의미있는 패턴을 발견하고 주로 상관성에 기반하여 데이터를 분석하고 새로운 통찰을 하자고 주장합니다. 발표자께서 언급하고 있듯이, 구속사유인 도주 및 증거인멸이라는 기준은 구체화되어 있지 않습니다. 사실 구속 여부를 결정함에 있어 도주 및 증거인멸이라는 기준이 실질적인 역할을 하고 있는지 자체도 의문입니다. 오히려 범죄의 중대성, 재범의 위험성, 피해자 및 참고인에 대한 위해 우려 등이라는 구속요인이 구속 여부를 결정짓는 보다 본질적인 요소라고 생각됩니다. 그렇지만 판사들이 구속 여부의 결정 사유로 구속영장 결정문에 기재하는 것은 통상적으로 도주 및 증거인멸 우려 여부 등입니다. 그렇기 때문에 판사들이 구속영장 결정문에 구속 사유로 기재한 내용을 토대로 구속결정에 영향을 주는 의미있는 패턴을 발견하기는 현실적으로 곤란하지 않을까 생각됩니다.

그러한 이유로 구속결정에 영향을 주는 의미있는 패턴을 발견하기 위해서는 오히려 양형사유를 들여다보는 것이 더 바람직할 여지가 있을 것으로 생각됩니다. 구속결정에 영향을 미치는 가장 본질적이고 중요한 요소가 범행 자체라는 점에 대해서는 의문을 제기하기 어려울 것이고, 판결에 적시된 양형사유에는 이러한 범행 자체에 대한 평가가 가장 구체적이고 자세하게 나타날 것이기 때문입니다. 또한 양형사유에 대해서는 양형기준의 제정과 운용과정에서 많은 논의와 연구성과들이 축적되어 있기도 합니다.

다만 발표자가 확인하고 재정립하려는 대상이 구속사유임에도 이를 직접 들여다보지 못하고 양형사유에 의해 연구를 수행하는 것은 방법론상 부적절할 것이어서 하나의 아이디어 차원에서 위와 같은 의견을 제시합니다.

발표자는 “구속업무에 구속영장의 기각률이 높아지고 있는데 법원과 검찰간에 영장발부기준과 관련하여 의견이 일치되지 않아 불필요한 인력과 시간의 낭비요인이 되고 있는 것이다”라고 서술하고 있습니다. 구속영장의 기각률은 높아질 때도 있고 낮아질 때도 있을 것이므로 구속영장의

기각률이 높아지고 있는 특정한 시점에서의 사정을 일반화하여 구속사유를 분석하는 것은 부적절하지 않을까 생각됩니다. 그렇지 않더라도 오해의 여지를 없애기 위해 이 부분에 대해서는 부가적인 설명을 할 필요가 있어 보입니다.

형사법을 연구하는 사람들은 대부분 AI의 알고리듬을 제대로 알고 있지 못할 것이므로 이를 보다 상세히 소개해준다면 좋을 것 같습니다.

토론자의 역할에 충실하기 위해서 제대로 알지 못하는 내용에 대해 표피적이나마 비판적인 관점에서 의견을 개진해보았습니다. 인공지능에 의한 구속기준이라는 규범적인 영역의 분석이라는 미지의 영역을 개척하려는 발표자의 노력에 많은 응원을 보냅니다.

한국형사정책학회 춘계학술대회. 2022.3.25. 전북대학교

동국대학교 dongguk university

성폭력 피해자 조사용 인공지능 챗봇 개발과 형사정책적 함의

조은경 양기주

동국대학교 경찰행정학부 · 컴퓨터정보통신공학부

ekjo@dongguk.edu

qiyang@dongguk.edu

요약문

대면조사가 어려운 성폭력 피해자로부터 양질의 진술을 획득하고 2차 피해 가능성을 최소화하는 챗봇 개발이 최근에 주목받고 있다. 이 발표에서는 성폭력 피해 아동 면담 조사에 사용되고 있는 NICHD 면담기법의 원리를 구현하는 인공지능 챗봇을 개발하기 위해 자연어처리(NLP) 목적 지향 대화 시스템(Task-Oriented Dialogue System)을 적용한 국내 연구 사례를 소개하고 형사정책적 함의를 논의하고자 한다.

동국대학교 dongguk university

목차

1. 성폭력 피해자 조사의 어려움
2. 바람직한 성폭력 피해자 조사 방법
3. 성폭력 피해자 조사용 챗봇 개발의 효용성
4. 성폭력 피해자 조사용 인공지능 챗봇 개발 사례
5. 인공지능 챗봇의 형사정책적 함의

동국대학교 dongguk university

1. 성폭력 피해자 조사의 어려움

•성폭력 범죄에 대한 형사절차는 '무슨 범죄가 일어났느냐' 설명과 그 증거로서 범죄 피해자의 진술 확보가 출발점이다(황은영, 2010).

•물적 증거 없는 성폭력 사건들, 피해자 진술이 유일한 증거인 경우 많음
•법정에서 피해자 진술의 신빙성 문제가 유무죄 판단의 근거인 경우 많음
•피해자는 수사기관에서 무엇을 얼마나 많이 말해야 하는지 모르고 조사자의 질문에만 답하는 경향
•피해자와 라포형성 중요: 두려움, 수치심 등을 완화하고 진술동기를 촉진
•면담방법과 사용되는 질문의 유형에 따라 진술의 양과 질이 달라짐
•유도질문, 폐쇄형 질문은 진술에 왜곡을 초래할 가능성이 높아 지양
•면담자의 태도와 질문에 의한 2차 피해 초래 가능성

2. 바람직한 성폭력 피해자 조사 방법

- NICHD 프로토콜 (Orbach et al., 2000)
- (National Institute of Child Health and Human Development)
- 심리학 연구 결과를 바탕으로 아동 목격자의 발달 수준을 고려하여 개발된 면담 프로토콜
- 경험적 연구를 통해 여러 나라에서 효용성이 입증됨
  - (영국 Lamb et al., 2009; 미국 Sternberg et al, 2001; 이스라엘 Orbach et al, 2000; 캐나다 Cyr & Lamb, 2009; 한국 Yi, Jo & Lamb, 2015 등).
  - 조사의 구조화, 질문의 질적 향상, 아동 제공 정보의 양과 질 향상
  - 개방형 질문 사용 5배 이상 증가, 선택형 질문 50% 감소
- 미국, 영국, 이스라엘, 캐나다, 한국, 일본 등 여러 나라의 경찰 수사 실무에서 사용되고 있음
- 15개 이상 언어로 번역됨 (https://nichdprotocol.com/)

※ 성폭력 피해자 표준조사 모델 (경찰청, 2018)

THE NICHD PROTOCOL의 구조

소개와 기본 규칙들
↓
라포 형성
↓
일화적 기억 훈련
자유 회상
단서제시 질문
단서제시 진술권유
시간 구분하기
↓
사건 내용 회상

사건 내용 단계
피해 주장이 있는가?
기본적인 진술권유
"한 번인가 아니면 여러 번인가?"
각 사건에 대하여 :
단서제시 회상 질문
지시형 질문
말하지 않은 정보에 관한 선택형 질문
↓
폭로 단계
↓
종료
중립적 주제

조사자의 질문 유형 분류

질문 유형	정의	예문
진술권유	사건에 대해 아동이 자유롭게 회상할 수 있도록 하는 개방형 질문. 아동이 언급한 사항에 대한 세부 정보 자유회상을 촉진	"일어났던 일에 대해 모두 말해보세요", "그리고 무슨 일이 있었나요?"
구체적 질문	아동이 이미 언급한 정보에 대해 구체적인 추가정보를 요구. WH질문	"그 셔츠의 색깔이 뭐였나요?", "그가 어디를 만졌나요?"
선택형 질문	아동이 언급하지 않은 정보에 대한 확인 혹은 선택적 질문	"너를 때렸니 안 때렸나요?" "바지를 입었니 치마를 입었나요?"
암시형 질문 (유도질문)	아동이 언급하지 않은 정보를 가정하거나 암시하는 질문	"그 사람이 너를 아프게 했나요?" "억지로 그렇게 했군요, 그렇죠?"
촉진어	아동이 계속 말을 하도록 비암시적으로 격려하거나 아동의 말을 반복하는 것	"그래요", "응" "그렇게 했군요"

조사자 : 오늘 무슨 일로 여기에 오게 되었나요? (구체적 질문)
아 동 : 어떤 아저씨가 맛있는거 사준다고 오라고 했는데요
근데 저는요 엄마가요 집에 맛있는거 만들어준다고 했어요 안간다고 팔빼고 가는데요
계속 따라 왔어요
조사자 : 계속 얘기해봐요. (진술권유)
아 동: 그때요 신호등을 건너 놓고부터 따라 왔었어요
조사자 : 어디 신호등? (구체적 질문)
아 동: xx주공 앞에 있는 OO어린이집 있던 그 앞에
조사자 : 음 따라오더니 그 다음에 무슨 일이 있었는지 이야기 해줄래요? (진술권유)
아 동: 갑자기요 아파트로 1층 안에서요 뭐 뭐 아까 했던 말이랑요...
조사자 : 아저씨가 1층에 들어온 다음에 일어난 일에 대해 좀 더 자세하게 말해주세요. (진술권유)
아 동 : 맛있는 거 사준다고 하면서 갑자기 제 팔을 잡아서 끌어당기는 거에요. 그래서 제가...

3. 성폭력 피해자 조사용 챗봇 개발의 효용성

- 코로나19 확산으로 인한 비대면 활동 증가
- 물리적 제약 극복 : 지리적, 시간적 제약으로 인한 비대면 조사 가능
- 심리적 제약 극복 : 외상후스트레스, 보복의 두려움 등으로 대면 면담을 꺼려하는 피해자에게 심리적 부담 완화
- 대면 조사가 어려운 취약한 피해자로부터 양질의 진술을 획득할 수 있고, 2차 피해의 가능성 최소화
- 항시 조사 가능성: 웹사이트나 앱은 24시간 운용될 수 있어 수사면담에 대한 접근성을 높임

- 한국일보, 'AI 챗봇'이 성범죄 피해 진술받아 '2차 피해' 막는다, 2018.08.02., https://www.hankookilbo.com/News/Read/201808021348769510.

4. 성폭력 피해자 조사용 인공지능 챗봇 개발 사례

성폭력 피해자 조사용 챗봇 개발 단계

- 대검찰청 심리분석실 R&D 과제 (2020-2023)
- 성폭력 피해자 진술 데이터베이스 구축과
- 자연어 원시 데이터를 AI가 학습할 수 있는 형태로 가공하기 위해
- 대검찰청 심리분석실에서 조사한 성폭력 피해자 면담 녹취록 개인정보 삭제 작업
- 성폭력범죄 구성요건 분류체계 구축 (판결문 분석)
- 분류체계에 따른 녹취록 수작업 코딩, 약 300건
- 데이터 전처리
- NICHD 면담기법을 기반으로 스토리라인 구성
- 한국어 커스터마이징된 Rasa, Slot filling 방식
- 대화형 챗봇 구축 (진행 중)

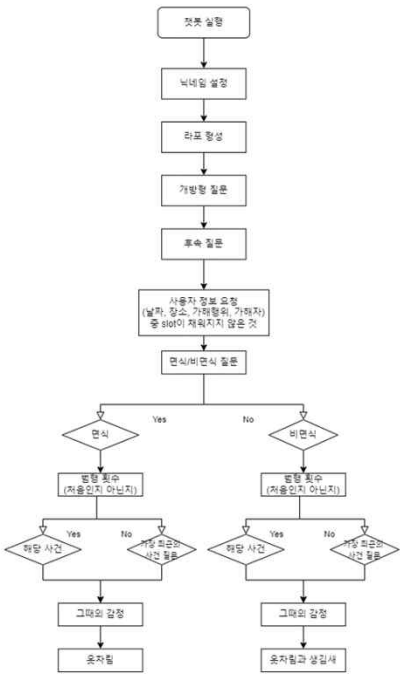
성폭력 피해자 조사용 챗봇과 일반적인 정보제공 챗봇의 차이

- 일반적인 챗봇은 사용자가 질문하면 챗봇이 데이터베이스 지식 기반 답을 제공하는 방식
- (예) 법무부 '버비': 법지식 제공 서비스

- 반면,
- 피해자 조사용 챗봇은 피해자가 진술을 입력하면 그 진술에 대해 더 구체적인 내용을 말하도록 이끌어가야 함.
- 피해자가 말한 내용 중 사건의 핵심내용, 부가적인 내용 등을 성폭력 피해자 진술 데이터베이스에서 확인하여 추가 탐색이 필요한 부분을 질문하여 채워나가는 방식

피해자 수사면담 기법을 적용한 챗봇 구성 원리

- 본격적 면담 전 피해자와 라포 형성 대화
- 개방형 질문으로 피해 사실에 대한 면담 시작
- 피해자의 대답을 분석한 결과에 따라 다른 구성요소 질문 제기
- 사건 구성요건에 부합되는 진술 확보의 정도에 따라 알맞은 질문 제기
- 더 이상 채워질 부분이 없을 때 면담 종료

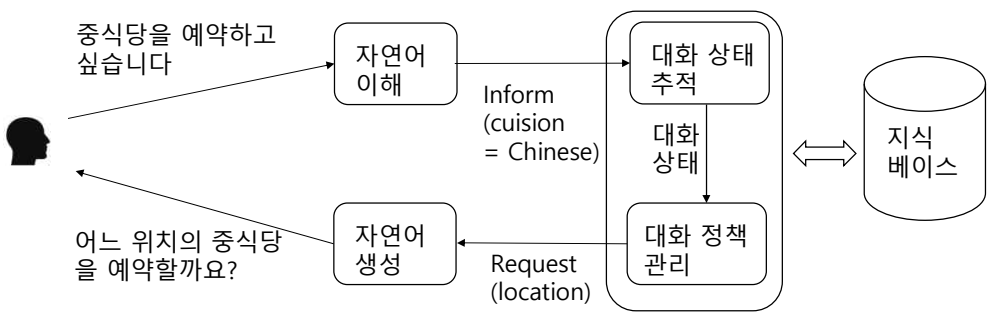


데이터베이스 구축: 문답 형태의 데이터 예시

문	지금부터는 00님께서 오늘 저에게 해주실 이야기를 처음부터 끝까지 모두 이야기 해주세요.
답	2학년이나 3학년 때 애들이 모르잖아요. 아예 그런 관련 자체를 그때는 성교육을 받았던 때도 아니니까요. 그래서 저는 정말 아무것도 모르고 학교에서 영화나 그런 곳에서 그런 성적인 영화가 나오니까 제가 애들한테 듣고, "그게 뭐야? 그게 뭐야?" 그러다가 [놀이터](장소)에 가서 [박지원 씨](attacker)에게 물어봤죠. 그런데 자기가 [가르쳐주겠다고 저한테 그런 행동을 해서](incedent) 당황했어요
문	그 이후에 있었던 일을 전부다 얘기해주세요
답	집에 들어왔을 때 제가 누웠거든요, 방이 너무 더러워서 방 먼저 다 치워놓고 누웠는데 [제 옷을 벗기고](nextincedent) 박지원씨도 자기가 직접 벗고 그냥 몇 번 하다가 [제 옷에다 쌌어요](nextincedent).
문	언제 그런 일이 있었나요?
답	[3월 25일](detail_date)이요
문	3월 25일이라고 하셨는데 그 날이라고 생각한 이유가 있을까요?
답	그냥 반팔, 반바지를 입었던 걸로 기억나요
문	박지원씨를 만난 적이 있나요?
답	네 (또는 아니오)
문	말씀해 주셔서 감사합니다. 이런 일이 혹시 한번 있었나요, 여러 번 있었나요?
답	[처음](횟수)이요 (또는 두번째요, 여러 번이요 등등)

면담 챗봇 시스템 구조

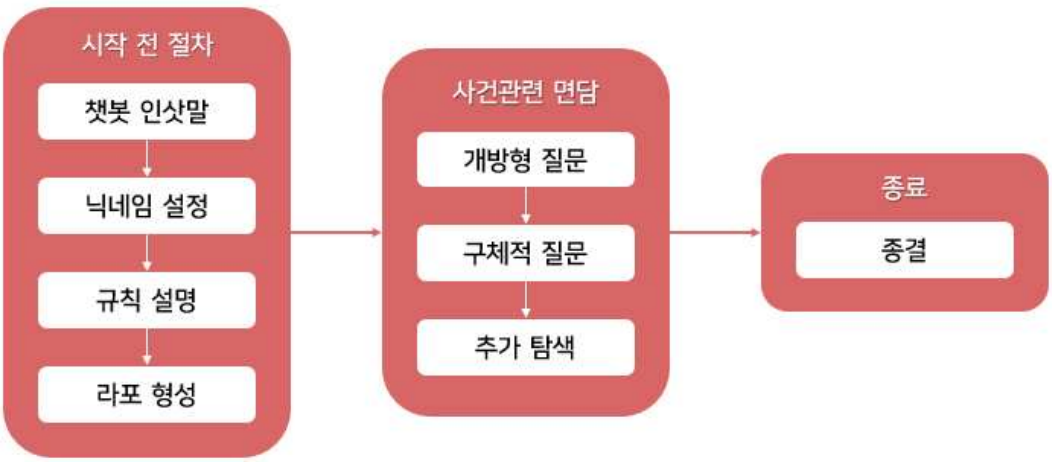
- Task-oriented Dialogue System (ToDS)



Rasa – open source platform for ToDS

- 정보입력 : 사용자 진술 입력
- Slot filling : 입력된 진술을 자연어 처리 후 entity 추출 및 주요 내용의 slot filling
- Intent tracking : 대화 이력 및 진술과 질문의 내용에 따른 사용자 의도 파악
- 질문 생성용 policy learning : 파악된 의도와 대화 이력 및 진술의 내용에 기반하여 다음 action을 선택하고 response를 생성하는 과정을 학습

구현된 챗봇 프로세스



구현된 intent, entity, action, response

▲ Intent 종류

▲ Entity 종류

▲ Action 종류

▲ Response 종류

5. **인공지능 챗봇의 형사정책적 함의**

피해자 조사용 챗봇 개발의 한계

- 자연어 데이터베이스 구축의 어려움. 피해자 진술 데이터 확보 어려움
- 자연스러운 라포형성 어려움: 라포의 질적 수준
- 진술자의 의도를 반영한 후속 질문 추출의 기술적 난이도
- 적절한 질문유형 선택 어려움
- 충분한 진술이 이루어졌는지에 대한 기준 설정 모호함
- 진술 일관성 및 구체성 판단 알고리즘 개발 필요
- 서면 형태 진술(타이핑)의 피로도 유발 가능성

5. 인공지능 챗봇의 형사정책적 함의

- 챗봇을 통해 획득한 진술(텍스트)은 데이터로 즉시 저장되고, 곧 이어 진술분석 알고리즘에 입력(input)될 수 있으므로 신속하게 진술에 대한 분석 결과(output)를 얻을 수 있다는 효용성은 있으나,
- 챗봇이 사건의 핵심적인 구성요소에 해당하는 진술을 충분히 획득할 수 있는가?
- 범죄 구성요건과 판례 및 조사 시 고려되는 요소들에 대한 데이터베이스 구축이 가능한가?
  - 가해자와 피해자 인물 정보, 피해 당시 정황, 맥락 정보, 사건 전후 맥락, 정황 증거, 폭행/협박 여부 판단에 필요한 행위 정보 등을 구체화한 지식 데이터베이스 필요함
- 챗봇의 수행에 대해서 수사면담 전문가들이 지속적으로 피드백하고 이를 기술자들이 알고리즘에 반영하여 보완하는 과정이 필요함
- 챗봇이 조사자의 지위를 획득할 수 있겠는가?
- 챗봇에 의한 진술과 대면조사 진술이 불일치할 경우 피해자의 진술에 대한 판단은 어떻게 할 것인가?

< 토 론 문 >

“성폭력 피해자 조사용 인공지능 챗봇 개발과 형사정책적 함의”에 대한 토론문

안성조(제주대학교 법학전문대학원 교수)

성폭력 피해자의 두려움과 수치심 등을 완화하여 진술동기를 촉진시킬 수 있다는 장점을 갖고 있는 ‘성폭력 피해자 조사용 인공지능 챗봇(이하 성폭력 조사용 챗봇)’의 개발취지와 활용방법 및 형사정책적 의의와 한계에 대한 귀중한 공동연구를 잘 받아보았습니다. 소위 4차 산업혁명시대에 AI 기술의 법제도적 응용분야가 과연 어디까지인가 실감할 수 있는 자료라고 생각합니다.

저는 전문적으로 성폭력 조사용 챗봇의 연구에 천착해 본 바는 없습니다만, 학내의 인권센터에서 일하면서 성희롱·성폭력 사건을 수차례 조사, 처리한 바 있어서 상당히 관심이 많이 가는 발제였습니다. 그리고 마침 인공지능 챗봇 개발과 관련된 윤리적 문제에 대한 연구서를 어느 모임에서 접할 기회가 있었던 터라서 이 기회에 그 내용을 공유하는 취지에서 몇 자 질문 겸 코멘트를 드리고자 합니다.

우선 성폭력 조사용 챗봇의 효용성이 좋게 평가되고 있는 것은 AI 심리치료상담사에 대한 선호가 높은 것과 매우 유사하다고 생각됩니다. 관련 연구[1]에 의하면 오늘날 미국에서는 수천 명의 사람들이 AI 심리치료상담사(이하 AI 치료사)에게 고민거리를 털어놓는다고 합니다. AI 치료사는 비용을 줄여준다는 점에서 장점이 있기도 하지만 사람들이 이 방식을 선호하는 이유는 이 컴퓨터 치료사가 어떤 평가를 내리지 않으며(nonjudgemental), 많은 정보를 제공해주기 때문에 보다 더 ‘자유롭고’ ‘솔직하게’ 대화에 임하게 되어 심리치료 효과가 높기 때문이라고 (psychologically effective) 합니다. 하나 흥미롭고 주목할 점은, 만일 이러한 개방성(openess)과 정직성(honesty)을 높이기 위해서라면 인간치료상담사가 대면이 아닌 컴퓨터를 통해 비대면 상담을 하는 방식도 가능할 수 있는데, AI 치료사는 이러한 인간상담사의 비대면 상담보다 개방성과 정직성을 높이는 데 보다 더 효과적이라고 합니다. 그 이유는 인간은 사회적 존재이기 때문에 사람들은 타인과 대면하고 있다는 사실을 알게 되면 있는 그대로의 모습보다는 남들이 더 선호할 모습을 보이려고 덜 정직한(less honest) 진술을 하게 되기 때문이라고 합니다. 요컨대, 상대가 인격체라는 암시는 우리의 행동에 영향을 준다는 것입니다. 그런데 사람들은 AI 치료사가 “인공적이고 의식이 없다는 것을 알고 있고”, 그래서 “비밀을 털어놓는 데 편안함을 느끼면서”, 또한 동시에 “이러한 인식을 접어 두고 이 프로그램이 의식이 있고 공감해주는 존재인 것처럼” 대함으로써 인간상담사에 비해 더 자유롭고 솔직한 상담이 가능해진다는 것입니다.

추측컨대, 성폭력 조사용 챗봇이 상당히 긍정적인 성과를 보이고 있는 것은 바로 이와 유사한 심리적 기제가 작동하기 때문이라고 봅니다. 기술의 발달로 자연계에 존재하지 않던 외견상 쾌고

1) Judith Donath, Ethical Issues in Our Relationship with Artificial Entities, in: The Oxford Handbook of Ethics of AI (Oxford Univ. Press, 2020) 참조. 이하 AI 챗봇의 윤리적 문제점은 이 책의 내용을 발췌, 요약, 번역한 것임을 밝혀둡니다.

감수성 있는 인공체(seemingly sentient artificial entity)가 새롭게 등장함으로써 인간과 상호작용하여 더 만족스러운 치료나 상담의 효과를 가져올 수 있다면 일단 피해자(환자)와 조사자(상담자)의 관계형성 측면에서나 형사정책적인 측면에서, 도구적으로 볼 때 긍정적평가가 내려질 수 있음은 분명하다고 봅니다. 그런데 여기서 한 가지 짚고 넘어가야 할 측면이 있습니다. 이 문제는 AI 윤리연구자들이 지적하는 것들입니다.

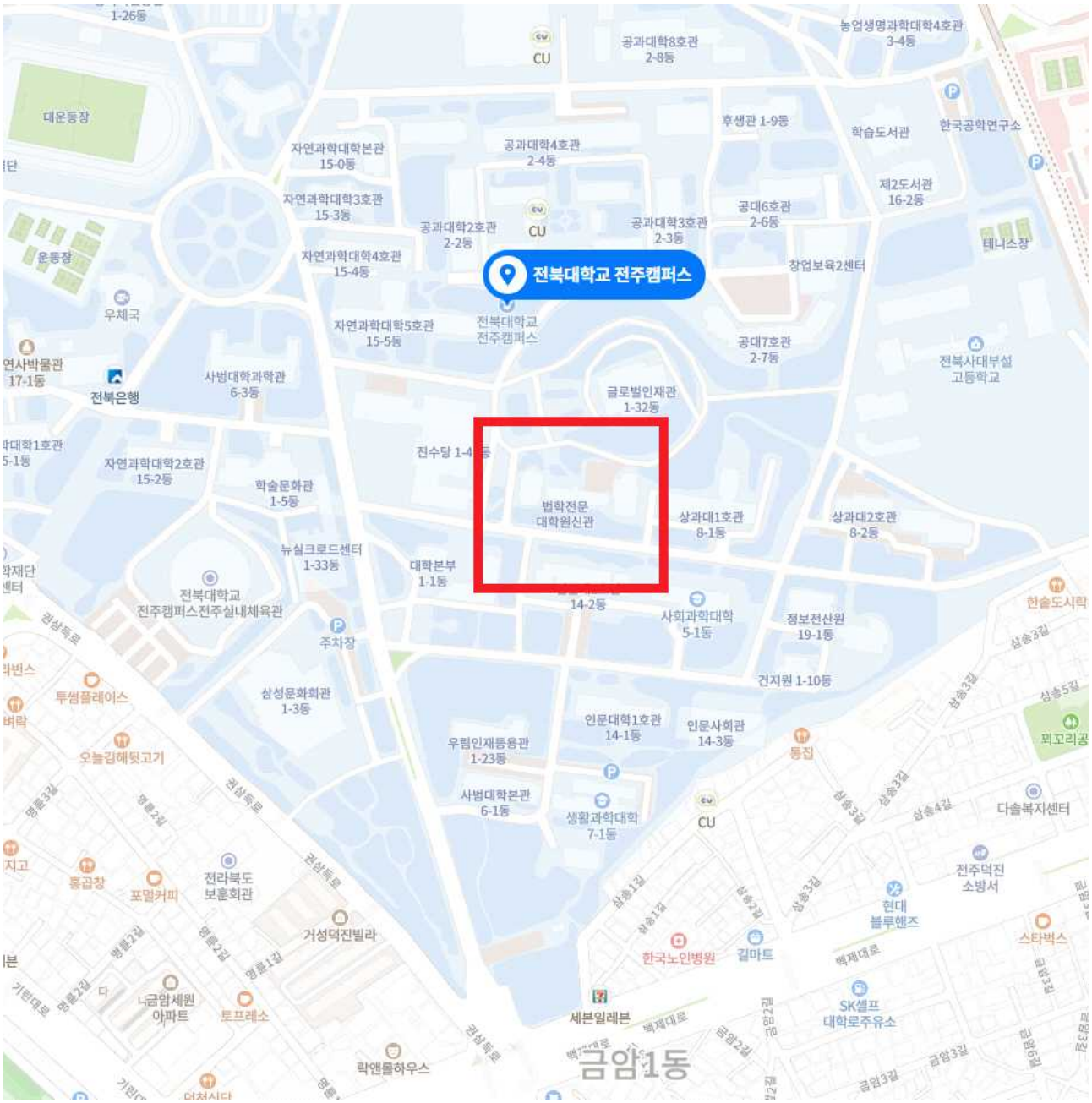
그것은 환자들이 성폭력 피해자들이 AI 챗봇을 선호하고 심리적으로 의지하는 것이 모종의 기망(deception)의 힘에 기초하고 있다는 사실입니다. 다시 말해 실제로는 의식도 감정도 없지만 외견상 그러한 존재처럼 보이는 존재로부터 오는 특수한 감정과 신뢰는 그 누구도 의도하지 않았지만(때로는 선한 목적에서 의도할 수도 있겠지만), 환자나 성폭력 피해자들이 AI를 인간과는 다른 모종의 인격체로 생각하며 대화를 하면서 자신의 생각을 털어놓게 하는 것은 일종의 기망이며, 따라서 그로부터 얻어내는 진술은 진술자의 자율성(autonomy)을 침해할 수 있다는 것입니다. 예컨대, AI 치료사가 상담을 시작하면서 “나는 AI 치료사입니다”라고 말한다면 그 자체가 기망적일 수 있는데, ‘나’라는 표현 자체가 사고하며 자의식이 있는 존재가 스스로 사고과정을 통해 이러한 단어를 형성했음을 암시하는데, 이는 ‘정체성 기망(identity deception)’이 되기 때문입니다. 성폭력 조사용 챗봇에서 설령 이러한 맹점이 모두 제거된다고 하더라도 AI 상담사가 기술적으로 점점 더 개선된다는 것은 더욱더 ‘인간과 유사한’ 질문을 하는 수준에 도달한다는 것이므로 이 역시 선한 목적으로 정체성을 속인다는 범주에서 크게 벗어나지 않을 것으로 보며 그것이 과연 윤리적으로 허용되는 것인지 고민이 좀 더 필요할 것 같습니다. 자율성의 원칙에 의하면 조작되지 않은 채 자신의 결정을 내릴 능력을 빼앗는 것은 그 자체로 비윤리적인 것이 될 수 있기 때문입니다.

또 다른 기망의 문제가 있습니다. 그것은 프라이버시(privacy)와 관련된 것입니다. 환자나 성폭력 피해자는 자신이 AI 상담사(치료사)와 대화함으로써 보다 자유로우며 비밀이 보장된다고 느낄 수 있지만, 실제로 그 대화내용은 저장되고 다른 사람들에 의해 읽히고 분석되고 있음을 알게 되면 매우 불편할 것입니다. 이러한 프라이버시 침해의 문제를 사전에 충분히 고지하고 성폭력 조사에 임한다면 과연 원래 기대했던 만큼의 효과를 얻을 수 있을지 조금 더 면밀히 검토된다면 형사정책적으로 바람직한 챗봇개발 및 그 윤리적 운용방침의 수립에 도움이 될 수 있으리라고 기대합니다.

끝으로 성폭력 피해자 챗봇의 조사자 지위는 현행법의 개정이 없는 이상 인정되기 어려울 것으로 보이며, 다만 챗봇을 통한 피해자의 진술기록은 그것이 AI의 보조를 통해 피해사실 진술에 도움을 받은 정도라면 진술서(제312조 5항 또는 제313조)로서, 만일 그것 자체가 수사기관 작성의 참고인 진술조서의 보조자료로 활용된 경우라면 진술조서(제312조)로서 증거능력을 부여받을 수 있을 것으로 전망됩니다. 아울러 챗봇에 답변한 내용과 실제 조사자와 대면하여 답변한 내용이 상이할 경우, 이는 진술의 증명력 판단에 고려될 문제로서 법관의 자유심증(제308조)에 맡겨져 있는 것이므로 법적으로 큰 문제점을 초래하지는 않을 것으로 사료됩니다.

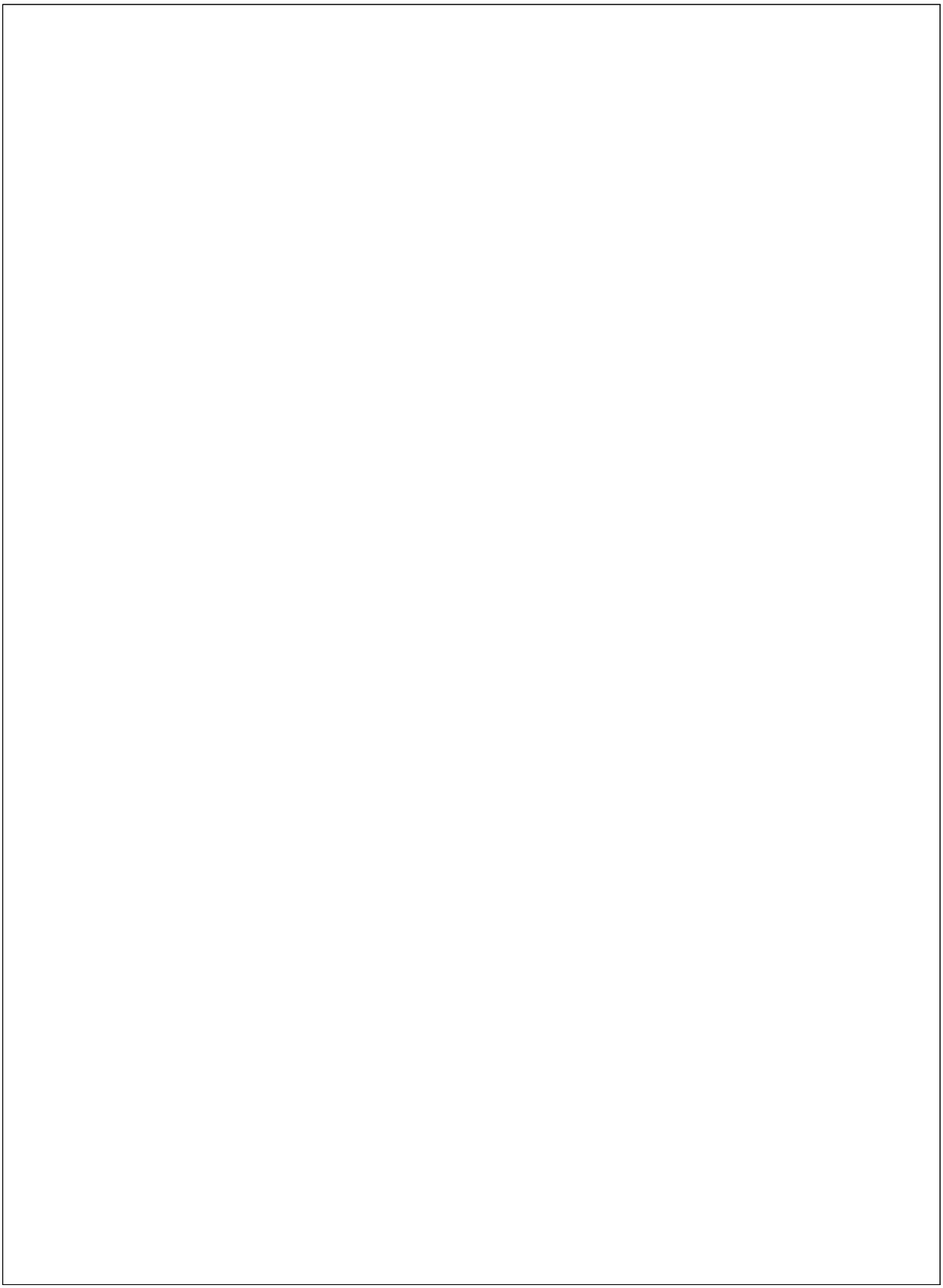
성폭력 피해자 조사에서 발생하는 매우 어려운 문제점을 AI 기술로 극복해 보려는 시도에 대해 함께 고민해 볼 수 있는 유익한 발제를 해 주셔서 감사합니다.

찾아오시는 길(Map)



○ 주소 : 전라북도 전주시 덕진구 백제대로 567(금암동 663),전북대학교 법학전문대학원

○ 자가용 : 호남고속도로 이용/ 전주 I.C부터 8.2km / 전주 톨게이트(우측방향) → 호남제일문 통과 → 팔달로 → 공설운동장 네거리(좌회전) → 전북은행 본점 → 전북대학교

○ 고속버스 : 전주고속터미널 또는 전주시외버스터미널에서 약 1.8km(택시 기본요금)

○ 기차 : 전주역에서 약 3.6.km (택시요금 약 5,000원)