규제연구 제32권 제2호 2023년 12월 http://dx.doi.org/10.22954/ksrs.2023.32.2.004

# AI 시스템의 위험 완화를 위한 정책적 접근방안 연구: AI 영향평가를 중심으로*

김 근 혜**· 박 규 동***

본 연구는 AI 시스템의 위험 완화를 위한 정책접근 방식으로 공공분야에서 발표한 AI 영향평가의 탐색적 사례분석 수행을 목표로 한다. 본 연구는 선행연구에서 제시한 AI 영향평가 설계의 핵심 사항을 토대로 연구분석 틀을 구성한다. 이후 분석 틀을 토대로 현재까지 국가 정부 혹은 공공기관에서 정책으로 구현하거나 제안한 6개의 AI 영향평가 사례를 분석하였다. 분석 결과 AI 영향평가는 기존 영향평가와 다른 기술적·산업적 특수성을 인정해야 한다는 인식이 존재했으며, AI 시스템을 포괄적이고 광범위하게 정의했다. 위험 완화조치와 관련하여 일부 의견이 수렴하고 있는 것을 확인하였지만, AI 영향평가에서 제시하는 완화조치가 위험분류 수준과 연결되는 방식이 국가마다 상이했다. 또한, AI 수명주기의 여러 지점에서 영향평가를 수행했으며 자율규제적 성격이 강했다. 본 연구는 AI 영향평가 개발과 관련하여 현재까지의 정책구현 동향을 살펴봄으로써 AI 영향평가 개발의 전반적 이해를 도울 수 있는 포괄적 자료를 제공하고, 추후 정책개발에 정보를 제공할 수 있다는 의의가 있다.

핵심 용어: 인공지능, 인공지능 영향평가, 알고리즘 영향평가, 자동화된 의사결정

---

* 이 논문은 2023년도 광운대학교 교내 학술연구비 지원에 의해 연구되었음
** 제1저자, 고려대학교 정보보호연구원 연구교수, 서울시 성북구 안암로 145 (stargazer@korea.ac.kr)
*** 교신저자, 광운대학교 행정학과 교수, 서울시 노원구 광운로 20 (gdpark29@kw.ac.kr)
접수일: 2023/12/23, 심사일: 2023/12/23, 게재확정일: 2023/12/26

# I. 서론

인공지능(Artificial Intelligence, 이하 AI)은 산업 전반에 걸쳐 혁신을 가속했으며, 그 과정에서 비즈니스 방식을 재창조하며 점점 더 다양한 애플리케이션과 시스템에 배포되고 있다. 전 세계 정부도 이민자 선별, 복지 자격 판단 등 다양한 공공서비스에 의사결정을 자동화하거나 지원하기 위해 AI 시스템을 도입하고 있다. AI 기술이 공공기관과 산업 전반에 광범위하게 적용되고 사회에 미치는 영향이 커지면서 AI 규제는 점점 더 중요하고 복잡한 문제가 되었다. AI 도입 초기에는 AI의 개발과 사용을 주관하기 위한 규정보다는 발전과 성장, 잠재적 응용 분야 탐색에 중점을 두었다. 그러나 AI의 사용이 보편화되고 의존도가 높아짐에 따라 최근에는 AI의 잠재적 위험을 평가하고 관리하기 위한 규제 접근방식에도 관심이 커지고 있다. 전 세계 정부는 AI 시스템을 규제하기 위해 다양한 접근방식을 구현하고 있다. 몇몇 정부는 AI의 개발과 사용을 감독하기 위해 윤리적 프레임워크를 만들고 규제 기관을 설립하였다. 일부는 AI를 사용하는 영역과 주체별로 AI 규제에 접근하는 방식을 취하고 있다. 국가마다 제도환경에 따라 AI 규제에 접근하는 방식은 다양하다. 그러나 전반적인 추세는 AI가 책임감 있고 신뢰할 수 있는 방식으로 개발하고 사용하는 것을 보장하기 위해 감독을 강화하고 지침을 수립하는 정책 방향으로 진행되고 있다. 이는 경제협력개발기구(Organisation for Economic Co-operation and Development, 이하 OECD)에서 AI 원칙(OECD Principles on AI) (2019)을 발표하고 많은 국가와 조직에서 이에 동조하는 원칙을 마련하면서(Ezeani 외, 2021) 더욱 강화되었다. 또한, 위험관리 관점의 AI 규제 거버넌스가 주류 접근방식 중 하나로 부상했다(OECD, 2023). 다양한 영역에서 행위자들은 책임감 있고 윤리적인 AI 관행 채택의 필요성을 강조하며 AI 시스템의 위험을 식별하고 영향을 평가하기 위한 다양한 정

책대안을 제시하고 있다. 주목할 점은, 많은 공공분야에서 AI 영향평가(Artificial intelligence Impact Assessment, 이하 AIA)[1]를 주요 규제 방안으로 권장하고 있다는 점이다. AIA는 AI 시스템 수명주기 전반에 발생 가능한 문제를 파악하고 적절한 완화 조치를 시행하여 알고리즘의 책임을 강화하는 정책 메커니즘으로 알려져 있다(Reisman, 2018). 최근 전 세계 공공기관, 산업기관, 표준화 기구, 시민단체 등에서 AIA를 수행하기 위한 다양한 유형과 지침을 개발했다. 이와 더불어, AI의 사용이 미치는 직접적/잠재적 영향을 연구하고 발생 가능한 문제를 해결하기 위한 조직과 이니셔티브를 설립했다. 그러나 국내에서는 AIA 사례에 대한 소개에 그칠 뿐 학술적 분석이나 검토는 아직 부족한 실정이다. 이에 본 연구는 최근까지 공공분야에서 발표한 AIA의 주요 내용과 범위를 분석하고 AIA 구현 동향을 파악하는 것을 목표로 한다. 특히, AI 시스템의 위험을 완화하고 알고리즘의 책임을 달성하려는 목표는 비슷하지만, 제도적 환경에 따라 다양하게 구현되고 있는 AIA 비교분석을 통해 AIA 프로세스에 대한 정책 추세를 확인하고자 한다.

# Ⅱ. 이론적 고찰

## 1. AIA의 개념 및 특징

AIA에 대한 합의된 정의는 없다. 기존 연구는 AIA는 AI 시스템의 설계, 개발, 배포과정에서 조직이나 사회에 발생할 수 있는 다양한 위험 수준을 식별하고, 이를 완화하거나 제거하기 위한 관리 도구 또는 평가 기준으로 설명한다(Sean 외, 2022; Chae, 2020; Kazim 외, 2021). AIA는 조직이 AI 시스템 사용 시 의사결정 과정을 문서로 만들고 작업수행 과정을 공개하도록 요구함으로써 책임을 강화한다. 또한, 조직이 AI 시스템 사용주기 모든 단계에서 이용자와 사회에 미치는 광범위한 영향을 이해하고 이를 고려할 것을 권장한다. 여기에는 경제, 사회, 문화, 환경을 포함한 AI의 잠재적인 영향뿐만 아니라 상대적 이점과 비용, 잠재적 위험에 대한 평가도 포함되므로 위험관리 도구가 되기도

1) 알고리즘 영향평가(Algorithmic Impact Assessment)로 부르기도 한다.

한다(Lomg 외, 2022).

현재 공공분야에서 영향평가를 주요 규제방식으로 채택하더라도 AIA가 AI 시스템을 가장 효과적으로 규제할 수 있는 것은 아니다. 현재 AIA 제도는 정책의 최종결정에 영향을 미칠 만큼 효과적이지 못하다. 그러나 AI 서비스의 사용에 있어 개발자, 정책 입안자, 대중과의 정보 격차가 존재한다는 점을 고려할 때, 개발 과정을 늦추고, 대중의 의견을 수렴할 수 있는 경로를 만들고, 정보를 공개하는 등의 규제를 통해 현재의 피해를 완화하고 미래에 더 효과적인 규제를 개발하기 위한 주요 과정이 될 수 있다(Selbst, 2021). AIA 규제환경에서 만들어지는 관련 정보와 문서작업 역시 AI 사용과정을 투명하고 책임감 있게 운영하는데 긍정적인 작용을 한다.

한편, AIA의 등장은 AI 윤리의 발전과도 깊게 연결되어 있다. AI와 자동화된 의사결정(Automated Decision System, 이하 ADS)의 윤리적 영향에 대한 논의는 이미 1950년대와 1960년대에 제시되었다(Samuel, 1959; Wiener, 1988; Lo, 2020 재인용). 그러나 최근 AI 시스템의 개발 및 배포가 증가하고 피해사례가 증가하면서 AI 사용에 대한 우려가 커지고, 사회적 영향과 윤리적 영향에 대한 인식이 다양한 이해관계자, 학계, 정부, 시민사회, 산업 내에서 증가했다. 이러한 결과로 AI 윤리, 신뢰할 수 있는 AI, 책임 있는 AI가 AI 연구에서 주류적 접근법으로 부상하며 이에 대한 논의가 활발히 이루어지게 되었다. Kazim 외(2021)는 AI 윤리가 세 단계를 거쳐 왔다고 설명한다. 첫 번째는 AI 윤리원칙 제정 및 배포 방식이다. 두 번째는 공학 중심의 문제 해결에 초점을 맞춘 윤리적 AI 설계 접근방식이다. 세 번째는, AI 윤리원칙을 표준화하고 구체화하여 운영하는 접근방식이다. 이에 따라 현재는 신뢰할 수 있는 AI, 책임감 있는 AI 사용 달성을 목표로 규제 거버넌스, 영향평가, 조달, 감사와 같은 다양한 프레임워크가 제안되고 공식화되는 지점까지 발전했다. 특히, AIA는 AI 윤리의 전문화를 향한 중요한 단계로 개인정보, 안전, 위험, 보안, 편향 등을 포함한 윤리적 모범 사례 적용과 관련하여 조직이 약속을 이행했는지를 확인하는 데 도움을 준다(Stahl 외, 2023).

최근 10년간 간 국가가 경쟁적으로 AI 전략을 발표할 때만 하더라도 AI 시스템 평가에 대한 용어 정의, 범위, 설계에 대한 합의는 거의 없었다. 많은 분야에서 AI 사용이 증가하면서 AI 시스템의 적합성과 평가에 대한 필요성에 대한 논의가 제기되었으며 영향평가의 역할이 두드러지기 시작한 것은 최근 5~6년 사이이다. 2018년 AI Now 연구소는

보고서에서 알고리즘 영향평가(Algorithmic Impact Assessments)라는 개념을 도입했다. 해당 보고서는 알고리즘 영향평가의 목적에 대해 지역사회와 이해관계자가 AI 시스템을 평가하고 사용의 적절성을 결정하는 데 도움을 주기 위함이라고 설명한다. 이후 학계를 중심으로 AIA 논의가 구체화 되었으며, 몇몇 정부에서 정책 형태로 소개되었다.

이러한 AIA는 새로운 정책개념이 아니다. AIA는 다양한 목적으로 광범위한 분야에서 사용하는 정책 평가 유형인 영향평가 방법론을 기반으로 한다. 일반적으로 영향평가는 해당 업무의 결과를 예측하기 어렵거나 사회에 미치는 영향을 측정하기 어려운 경우, 관련 업무자와 기술자가 영향평가를 할 수 있는 전문 지식을 가지고 있지만, 해당 정보를 문서로 작성하여 근거로 남길 유인책이 충분하지 않을 때 유용하다. 또한, 사회적, 경제적, 환경적 측면에서 특정 사업이나 정책이 미치는 영향을 사전에 심층적으로 분석하고, 이를 통해 미래에 발생할 수 있는 결과들에 대한 대처방안을 마련함으로써, 보다 효과적이고 합리적인 정책 결정을 돕는 중요한 분석 및 평가 도구로 그 가치가 인정된다(Selbst, 2021)[2]. 영향평가는 1969년 미국의 국가환경정책법(National Environmental Policy Act of 1969, 이하 NEPA)과 20세기 중반의 환경운동에 뿌리를 두고 있다(Selbst, 2021). NEPA가 제정된 이후 수십 년간 민간 부문에서 영향평가 방법론이 개발되었으며, 범위가 점차 확대되었다. 현재는 사회적 영향평가(Social Impact Assessment), 환경영향평가(Environmental Impact Assessment), 인권 영향평가(Human Right Impact Assessment, 이하 HRI)뿐만 아니라 데이터 보호 영향평가(Data Protection Impact Assessment, 이하 DPIA), 윤리 영향평가(Ethical Impact Assessment, 이하 EIA) 등 주제별 영향평가의 형태로 오랜 역사가 있으며. 지속 가능하며 정보에 기반한 의사결정을 내리는 데 중요한 역할을 하고 있다. 국내 공공기관에서는 환경부의 환경영향평가와 건강영향평가, 과학기술정보통신부의 기술영향평가,

2) 영향평가는 다양하게 정의된다. 유럽위원회(European Commission)(2017)는 영향평가를 정책 입안자가 정책 목표 달성을 위한 최선의 설계 방법을 모색하는데 필요한 정보를 제공하는 핵심 도구라고 설명한다. OECD(2020)는 영향평가가 정보에 기반하여 계획된 정책 수단의 비용, 결과, 부작용 등을 평가하는 활동으로 정책효과 향상을 위해 필요한 도구임을 강조하고 있다. 국내에서 노화준(1986)은 영향평가가 정책대안 혹은 사업이 특정 집단이나 커뮤니티에 미칠 수 있는 환경적 또는 사회경제적 영향을 파악하고 예측하여, 정책 결정 과정에 도움이 될 수 있는 중요한 정보를 제공하는 과정으로, 정책의 결과를 분석하고 평가하는 데 사용된다고 정의한 바 있다.

국가인권위원회의 인권 영향평가 등 다수의 영향평가가 운영되고 있다.

AIA와 밀접하게 관련된 영향평가는 DPIA가 있으며 HRI에서 정책 평가 부분과 데이터 권리 기술 관련 내용을 일부 참고하고 있다(Kaminski 외, 2019; Stahl, 2023). 기존의 영향평가와 달리 AIA는 초기 단계로 아직 AI 시스템을 평가할 수 있는 일관된 구조를 제공하는 방법론이나 국제적으로 승인된 표준은 없다. 그러나 기존 연구에서는 AIA 유형을 분류하는 작업을 수행하고 있다. 기존 연구에서 소개한 AIA 모델은 크게 4가지이다. 첫째는 질문형 모델(The Questionnaire Model) (Selbst, 2021; 김근혜 외, 2022 재인용)이다. 질문형 모델은 공공부문 AI 애플리케이션을 위한 질문 및 답변 형식으로 캐나다에서 2019년 ADS 관련 지침에서 취한 접근방식이 반영된다(Government of Canada, 2019). 설문지 모델은 기존 다른 분야의 영향평가와 달리 간단하다. 이는 AI 기술혁신의 특성상 짧은 시간 내에 평가 결과를 필요로 한다는 점에서 큰 장점으로 평가받고 있다. 두 번째는 DPIA 모델(Selbst, 2021; 김근혜 외, 2022 재인용)이다. 기존 연구들은 개인에게 높은 위험을 초래할 가능성이 있는 데이터 처리를 위해 유럽연합(European Union, 이하 EU)의 개인정보 보호법(General Data Protection Regulation, 이하 GDPR)에서 의무화한 DPIA를 AIA의 모델로 확장하여 사용할 수 있을 것으로 보고 있다. 기존의 DPIA는 알고리즘 의사결정보다는 데이터 처리의 결과로 발생하는 위험으로부터 사람을 보호하는 것을 목표로 하지만, GDPR의 DPIA는 ADS에 초점을 맞추고 알고리즘에 대한 자세한 설명과 작업 범위를 요청한다. Kaminski 외(2021)는 연구에서 GDPR의 DPIA의 투명성 요구를 법적으로 명시하지 않았다는 점에서, 설명받을 권리가 제한적이라고 한계를 설명한다.

세 번째는 공공기관 모델(The Public Agency Model) (IFOW, 2021)이다. 공공기관 모델은 가장 포괄적인 AIA 모델로 알려져 있다. 해당 모델은 투명성 유지, 이해관계자 참여, 규정 준수를 통해 공공기관, 대중, 규제 기관의 책임 관계를 공식화하며 유럽의회 연구 서비스(European Parliamentary Research Service, 이하 EPRS)가 제안하였다. 해당 모델은 다양한 수준의 사전 위험평가 절차가 포함되어 있다는 것과 대중의 참여를 특징으로 한다. 이러한 대중의 참여는 영향을 받는 커뮤니티가 중심이 되며 AI 시스템, 혹은 알고리즘에 대한 정보 공개, 피드백 요청 등이 포함한다. 공공기관 모델은 적극적인 모니터링을 통해 AI 시스템 사용에 따른 영향과 피해를 살피고, 이에 대한 광범

위한 협의를 통해 평가 프로세스를 수립하는 이점이 있다. 설명 가능한 선제적 절차와 계층화된 접근방식 역시 장점으로 언급된다.

마지막으로 국가환경정책법 모델(The National Environmental Policy Act(이하 NEPA) Model) (Selbst, 2021; 김근혜 외, 2022 재인용)이다. 미국 국가환경정책법의 영향평가 모델을 기반으로 하는 NEPA 모델은 수백 페이지에 달하는 상세한 영향평가로 개방형 질문에 대한 구체적인 답변을 요구한다. EPRS가 제안한 공공기관 모델 이전에 제시된 AIA 모델은 모두 NEPA 모델을 기반으로 하고 있다(Selbst, 2021). NEPA 모델 역시 투명성과 대중의 참여를 특징으로 하고 있으며 민간 부문보다는 공공부문에 초점을 맞추어져 있다. 그러나 일부 전문가는 NEPA 모델의 복잡성, 서류작업 및 규정 준수 비용, 느린 허가 절차, 긴 영향평가 기간으로 인한 기술혁신 지연 등을 고려할 때 AIA의 모델로는 적합하지 않다고 평가한다(Thierer, 2023).

문헌 연구는 기존 영향평가[3]와 마찬가지로 AIA를 사전영향평가(AI Risk Assessment)와 사후영향평가(AI Impact Evaluation)로 구분하여 개념화하기도 한다. 사전 평가는 AI 시스템 수명 주기 전반에 걸친 잠재적 영향을 고려한다면 사후 평가는 AI 시스템의 활용이 시작된 후 기술, 정책 또는 비즈니스 관행의 실제 영향을 고려한다(Long 외, 2022). Ada Lovelace Institute(2020)는 사전·사후 AIA를 〈표 1〉과 같이 구분한다.

3) OECD(2020)에 따르면, 공공 정책에 영향평가를 적용하는 두 가지 일반적인 방법은 다음과 같다. 첫째, 사전영향평가(Ex-Ante)로 정책 입안자들에게 미래와 현재 개입의 예상 효과를 알리기 위해 사전에 정책 주기의 요구분석 및 계획 활동의 일부가 된다. 둘째, 사후영향평가(Ex-Post)로 이전 정책 주기를 평가하고 관리한다. 사전 영향평가가 개입의 효과에 초점을 맞추지만, 사후 영향평가는 개입 설계의 적절성, 개입의 비용 및 효율성, 의도하지 않은 효과, 향후 개입의 설계 개선 등을 다룬다.

〈표 1〉 AIA 유형 분류

| 구분 | AI 사전 영향평가<br>(AI Risk Assessment) | AI 사후 영향평가<br>(AI impact Evaluation) |
|---|---|---|
| 정의 | (사용 전) AI 시스템이 사회적으로 미칠 수 있는 영향을 평가함 | (사용 후) AI 시스템이 사람에 미치는 사회적 영향을 평가함 |
| 사용 시기 | 배포 전 진행 | 배포 후 진행 |
| 사용자 | 개발자 혹은 위원(commissioners) | 연구원, 정책 입안자 |
| 기원 | 환경영향평가, 데이터 영향평가 | 일반적인 사후 영향평가 |
| 사례 | 캐나다 정부의 알고리즘 영향평가 | 스탠포드 대학교의 엘러게이니 카운티(Allegheny County) 아동 복지 사무소를 위한 예측 위험 모델링 도구 영향평가 |
| 상태 | 적용 가능성과 모범 사례에 대한 증거가 필요 | |

출저: Ada Lovelace Institute (2020)을 참고하여 재구성

# Ⅲ. 연구설계

## 1. 선행연구 및 연구 질문

AIA가 공공기관에서 활용하는 새로운 정책 메커니즘으로 주목받으면서((Ada Lovelace Institute 외, 2021) AIA의 학술적 관심은 서구권을 중심으로 증가해왔다. 선행연구의 주요 방법론은 질적연구, 탐색 연구, 이론분석 연구이다(김근혜 외 2022). 국외에서 AIA 관련 연구는 크게 4가지 연구 경향을 보인다. 첫째, 기존의 영향평가를 토대로 AIA에 대한 학술적 개념화를 시도한다. Mantelero(2018), Raab(2020) Yam 외(2021), Kazim 외(2021)는 DPIA와 EIA와의 비교를 통해 AIA의 주요 범위와 정의, 차이점을 분석한다. 둘째, AIA 관련 법률의 내용 분석이 주를 이루고 있다. Nahmias 외(2020), Kaminski 외(2019), Chae(2020), Kasirzadeh 외(2021)는 2018년 시행된 EU의 GDPR의 제35조(고위험 개인정보 처리에 대하여 개인정보영향평가를 의무화)와 2019년 미국 의회에서 제시된 알고리즘 책임법(Algorithmic Accountability Act

of 2019)을 분석함으로써 최근 AI 감독 메커니즘의 경향을 살피고 법률에서 제시된 각 AIA 제언의 한계와 개선점을 제시하고 있다. 세 번째, AIA의 실효적 도입방안을 위해 필요한 제도적 개선사항과 평가 프레임워크를 제안한다. Selbst(2021)는 AIA 규제가 민간분야와의 협력을 통해 제도권 내에서 효과적으로 도입되기 위한 다양한 방안을 살펴본다. OECD(2021)는 OECD AI 원칙을 토대로 신뢰할 수 있는 AI 시스템을 구현하기 위한 도구와 관행을 비교하기 위한 프레임워크를 제안한다. Reisman 외(2018)는 공공부문을 대상으로 하는 알고리즘 영향평가의 프레임워크와 핵심 요소를 제시하고 있다. Yeung(2021)은 AI 위험 및 영향평가 개발에 있어 필요한 핵심적인 권장 사항을 제시한다. 넷째, Stahl 외(2023)는 문헌분석의 체계적 검토를 통해 잠재적 사용자가 AIA 개발 시 필요한 적절한 요구사항을 제공한다.

국내에서 AIA 관련 연구는 지능정보화 기본법을 중심으로 사회적 영향평가에 관한 연구가 주를 이루고 있다. 유순덕(2023)은 인공지능 서비스의 영향성을 평가하기 위한 분석기준을 영향력, 지속가능성, 효율성, 효과성, 적절성으로 분류하여 제시하고 있다. 김법연(2023)은 국내 공공분야에서 AI를 도입 및 활용 시 안정적으로 AIA를 운영할 수 있는 제도개선 방안을 제안하고 있다. 권은정(2023a, 2023b)은 지능정보화기본법을 중심으로 사회적 영향평가 제도화 방안과 안정성과 신뢰성을 목표로 하는 방법론을 제시하고 있다. 김근혜 외(2022)는 문헌 검토와 내용 분석을 통해 AIA 관련 최근까지의 연구 동향을 분석했다.

그러나 국내외 AIA 관련 기존 연구들은 최근 제도권에서 AIA를 도입하기 위한 다양한 시도에 대하여 학술적 토대에 기반한 체계적 논의가 부족한 실정이다. 본 연구는 AIA에 대한 학술적 관심이 커지고 정책구현으로도 활발히 이루어지고 있음에도 두 사이의 틈을 확인하기 위한 체계적인 시도가 부족하다는 점에 주목한다. 현재 AIA에 대한 정책 도입은 초기 단계로 일관된 방법론이나 합의된 접근방식이 없다. 이러한 다양한 접근방식은 서로 다른 이해관계자를 대상으로 하고 있으며 서로 다른 이해관계와 목표 아래 개발되고 있다. AIA가 구성되고 구현되는 방식에도 상당한 차이가 있다. 본 연구는 AI 시스템의 위험 완화를 위한 정책접근 방식으로 2023년 현재까지 공공분야에서 발표한 AIA에 대한 탐색적 사례분석 수행을 목표로 한다. 해당 목표를 구체화하는 연구 질문은 다음과 같다. 첫째, 공공분야에서 발표한 AIA의 공통점, 차이점, 주요 특징은 무엇인가? 둘째,

학술적으로 제안하는 AIA와 정책으로 구현된 AIA는 어떠한 차이가 있는가?

질문에 답하기 위해 본 연구는 선행연구에서 제시한 AIA 설계 시 핵심 고려사항을 토대로 연구분석 틀을 구성한다. 이후 분석 틀을 토대로 현재까지 국가 정부 혹은 공공기관에서 정책으로 구현되거나 제안된 AIA를 사례를 채택하여 비교 분석한다.

## 2. 분석 틀

AIA의 형태는 국가나 조직에 따라 다양하지만, 일반적으로 AI 시스템의 사용으로 인해 발생할 수 있는 내재적 위험과 잠재적 피해를 포함하여 AI 시스템의 개발자 또는 배포자가 고려해야 하는 일련의 요소를 나열한다. 본 연구는 선행연구에서 제시한 AIA 평가 요소 및 요구사항 등을 토대로 연구분석 틀을 구성하였으며, (1) AIA 기본구조, (2) 위험평가, (3) 영향평가, (4) 위험 완화로 구분한다.

(1) 기본구조에서는 크게 4가지를 살펴본다. 첫째, AIA에서 AI 시스템의 정의와 평가 영역을 확인한다(OECD, 2021; Stahl 외, 2023). 둘째, AI 생명주기에서 AIA가 개입하는 시기(예: AI 시스템 배포 전, 후)를 살펴본다. 기존 연구는 AIA의 개입 시기의 여부가 AIA를 설계하는 핵심 고려사항이라고 설명한다(Selbst, 2021). 셋째, AIA의 평가방식(예: 폐쇄형 질문, 개방형 질문, 그 외)을 확인한다. 선행연구는 AIA의 일반적인 형태가 응답을 구하는 설문지 또는 체크리스트 형태의 질의응답 방식이라고 설명한다(Mantelero, 2018; Kaminski 외, 2019; Raji 외, 2020; Gebru 외, 2020; Selbst, 2021). 마지막으로 AIA 규제의 형태를 확인한다. 이는 감시 및 감독, 책임 메커니즘과 관련되어 있으며 여기에는 자발적(자율규제), 의무적(법적 규제) 등이 포함한다(Selbst, 2021).

(2) 위험평가 항목에서는 위험의 수준을 분류하고 데이터와 시스템의 위험을 식별한다. 기존 연구는 AIA 개발 시 위험 수준을 설정하고, 위험 수준에 따라 구체적인 조치 여부를 결정하여 요구하는 것이 위험을 완화하기 위한 좋은 대안이라고 제안한다(PwC, 2021). 본 연구는 위험을 식별하기 위해 AIA가 사용하는 데이터의 품질을 평가한다. 여기에는 데이터 민감도, 적시성, 적절성 등이 포함된다. 또한, 시스템 설계에서 투명성, 설명 가능성, 해석 가능성을 포함하고 있는지 확인한다(Reisman 외, 2018; Yeung,

2021). 투명성(Transparency)은 시스템에 무슨 일이 발생했는지에 대한 질문에 답할 수 있다. 설명 가능성(Explainability)은 시스템에서 결정이 어떻게 내려졌는지에 대한 질문에 답할 수 있다. 해석 가능성(Interpretability)은 시스템에서 결정이 내려진 이유와 의미 또는 맥락에 관한 질문에 답할 수 있다(NIST, 2023).

(3) 영향평가에서는 AI 시스템의 사용으로 개인 혹은 커뮤니티에 미칠 수 있는 영향의 특성과 규모(발생빈도, 심각성)를 살펴본다. 영향의 특성에는 기본권을 포함한 법적 권리에 미치는 영향, 개인의 신체적/정신적 안녕(well-being)에 미치는 영향, 경제적 영향 등이 포함된다(Yeung, 2021).

마지막으로 (4) 위험 완화이다. 기존 연구에서 완화조치는 AIA의 필수 구성요소로 강조되고 있다. 여기에는 편향 테스트 및 데이터 관리 여부, 최종결정 시 사람의 필수적인 개입과 감독의 가능 여부, 다양한 방식의 외부자 검토 및 참여 여부, 정기적인 모니터링 및 평가, 기술적·제도적 완화조치 한계에 따른 관련 AI 시스템 배포 금지가 포함된다(Reisman 외, 2018; PwC, 2021; Selbst, 2021; Yeung, 2021; OECD, 2021; Stahl, 2023). 해당 내용을 포함하는 연구분석 틀은 다음과 같다.

〈표 2〉 연구분석 틀

| 대분류 | 소분류 | 분석사항 (예시) | 참고문헌 |
|---|---|---|---|
| 기본구조 | 평가시스템 | AI 시스템에 대한 정의 | OECD(2021) Stahl 외(2023) |
| | 평가영역 | 해당 AIA가 사용 가능한 영역의 범위 | |
| | 개입시기 | 설계, 배포 전, 배포 후, 전(全) 단계 | PwC(2021) Selbst(2021) |
| | 평가방식 | 개방형 질문, 폐쇄형 질문, 그 외 | |
| | 규제방식 | 자발적/의무적 | |
| 위험평가 | 위험 수준 분류 | 수준별, 사이클별, 점수별 | PwC(2021) |
| | 위험식별 | 데이터: AI 시스템에 사용되는 데이터의 민감도, 적절성, 적시성 등 | Yeung(2021) Reisman 외 (2018) |
| | | 시스템: 설명 가능성, 투명성, 해석 가능성 등 | |
| 영향평가 | 영향의 유형 | AI 시스템 사용이 개인에 미치는 영향의 종류 | Yeung(2021) |
| | 영향의 발생빈도, 심각성 | AI 시스템 사용으로 개인에 영향이 미치게 될 발생빈도: 자주 있음, 가끔 있음, 거의 없음<br>영향의 심각성 정도: 낮음, 중간, 높음 | |

| | | | |
|---|---|---|---|
| 위험완화 | 편향완화 조치 | 편향 테스트 및 데이터 관리 | Reisman 외 (2018) PwC(2021) Selbst(2021) Yeung(2021) OECD(2021) Stahl(2023) |
| | 사람의 감독/개입 | 최종결정 시 사람의 필수적 개입 | |
| | 외부자 개입 | 외부 이해관계자의 참여 및 검토 | |
| | | 관련 자체 평가 및 검토 프로세스에 대한 대중 공개, 지역사회와의 협력 | |
| | | 시정 또는 이의제기를 위한 채널 설정 | |
| | 모니터링 | 정기적/지속적 모니터링 | |
| | 사용금지 | 기술적·제도적 완화조치 한계에 따른 관련 AI 시스템 배포 금지 | |

본 연구는 현재까지 국가 정부 혹은 공공기관에서 정책으로 구현되거나 제안한 AIA 사례를 조사하여 분석하였다. 학술기관, 표준협회, 시민사회조직에서 발표한 공공분야를 대상으로 하는 AIA는 사례에서 제외하였다. 사례선정 기준은 다음과 같다. 첫째, AIA에 일관된 방법론이나 합의된 접근방식이 없다는 것을 전제하고 AI 위험 프레임워크, AI 영향평가, AI 위험관리 등 AI 거버넌스 맥락에서 밀접하게 관련된 사례를 포괄적으로 선별한다. 둘째, 첫 번째 기준에서 채택한 사례 중 국가 정부나 공공기관에서 AIA를 수행하기 위해 실질적이고 구체적인 지침을 제공하는 문서를 채택한다. 본 연구는 독일, EU, 뉴질랜드, 미국, 캐나다, 네덜란드 정부가 구현하거나 제안한 AIA 사례를 선정하여 분석한다.

〈표 3〉 분석대상 국가와 AIA

| 국가 | AIA |
|---|---|
| 독일 | 정보윤리위원회의 의견 (Data Ethics Commission, 2019) |
| 캐나다 | 자동화된 의사결정에 대한 지침 (Government of Canada, 2019) |
| 뉴질랜드 | 아오테아로아 뉴질랜드 알고리즘 헌장과 가능성 및 영향평가 (New Zealand Government, 2020) |
| 유럽 | AI에 대한 규제 프레임워크 제안 (European Commission, 2021) |
| 미국 | 인공지능 위험관리 프레임워크 (NIST, 2023) |
| 네덜란드 | 기본권 및 알고리즘 영향평가 (Government of the Netherlands, 2022) |

해당 사례의 개요는 다음과 같다. 먼저 독일의 데이터 윤리 위원회(Data Ethics Commission)는 위험 규제 접근방식을 적용하여 AI 시스템 전반을 포괄하는 설계자 중

심의 위험관리 프레임워크(Opinion of the Data Ethics Commission)를 제안하였다. 해당 프레임워크는 인간중심의 AI 설계를 목표로 한다. 또한, AI 시스템의 중요성에 따라 규제 정도를 조정하는데, 시스템의 중요성은 시스템의 사용으로 발생할 수 있는 피해의 가능성과 심각성을 바탕으로 평가한다.

EU의 AI에 대한 규제 프레임워크 제안(Regulatory Framework Proposal on Artificial Intelligence)은 안전하고 윤리적이며 지속 가능한 AI 개발 및 사용을 보장하기 위해 제안된 일련의 법적 규정으로 유럽위원회(European Commission)가 2021년 개발했다. 해당 프레임워크는 위험 수준을 분류하고 있으며, AI 시스템의 개발, 설계 구현 시 적용하는 것이 가능하다. 또한, 정책 입안자, 비영리 조직, IT 부문 조직, AI 개발자와 공급자 모두를 대상으로 하고 있다.

뉴질랜드 정부는 공공기관에서 AI 알고리즘 사용으로 의도치 않게 악영향이 발생할 수 있는 위험을 완화하기 위해 2020년 7월 아오테아로아 뉴질랜드 알고리즘 헌장(Algorithm Charter for Aotearoa)을 발표했다. 헌장은 공공서비스에 사용하는 AI 알고리즘의 위험 발생 가능성을 자체적으로 영향평가(Assessing Likelihood and Impact)하고 평가 결과에 따른 위험성 등급에 따라 헌장의 적용 여부를 결정한다. 2023년 현재까지 27개 공공기관이 헌장에 서약했다.

미국 국립표준기술원(National Institute of Standards and Technology, 이하 NIST)은 2020년에 발표한 국가 인공지능 이니셔티브법(National Artificial Intelligence Initiative Act of 2020, PL 116-283)에 근거하여 AI 위험관리 프레임워크 로드맵(Roadmap for the NIST Artificial Intelligence Risk Management Framework 1.0, 2023)을 발표했다. NIST는 해당 프레임워크는 자발적이고 권리를 보호하며 특정 사용 사례에 구애받지 않고 모든 분야, 모든 규모, 사회 전체 조직에 프레임워크의 접근방식을 구현할 수 있는 유연성을 제공하기 위한 것이라고 목표를 밝히고 있다.

캐나다는 2019년 자동화된 의사결정에 관한 지침(Directive on Automated Decision-Making)을 발표한 후 이를 계속 업데이트하고 있다. 해당 지침은 조직이 ADS 사용과 관련하여 위험을 완화할 수 있도록 설계한 애플리케이션에 적합한 거버넌스, 감독 및 보고/감사 요구사항을 제공한다.

네덜란드 정부는 2021년 공공기관 내에서 알고리즘 사용을 위해 의무적 시행을 시행한 영향평가 프레임워크(Fundamental Rights and Algorithms Impact Assessment, 이하 FRAIA)를 도입했다. 해당 프레임워크는 알고리즘 개발, 조달, 사용을 고려하는 정부 조직을 대상으로 하는 토론 및 의사결정 도구로 결과가 아직 명확하지 않은 상태에서 알고리즘이 배포되는 것을 방지하기 위함을 목표로 하고 있다. 해당 프레임워크의 프로세스는 기본권 침해 위험을 중심으로 알고리즘 사용 시발생할 수 있는 결과(부정확성, 비효율성 포함)에 대한 질의응답을 포함한다.

# Ⅳ. 분석 결과

## 1. 기본구조

〈표 4〉는 본 연구에서 사례로 채택한 AIA의 기본구조를 분석하고 있는 것으로 평가시스템, 평가영역, 개입시기, 평가방식, 규제방식을 설명하고 있다. 먼저, 평가시스템과 관련하여 독일과 네덜란드는 알고리즘 시스템, EU는 AI 시스템, 뉴질랜드는 AI 알고리즘으로 정의하고 있으며, 미국과 캐나다는 평가 대상을 ADS로 시스템을 정의하고 있다. 이는 알고리즘 시스템의 특정 기능을 크게 제한하지 않고 모든 유형의 AI 시스템 혹은 알고리즘에 대하여 위험과 영향에 대한 평가를 고려한다는 것을 의미한다.

평가영역의 경우 뉴질랜드, 네덜란드는 공공서비스로 제한하고 있으나 독일, 캐나다, EU, 미국은 사회 전반의 사용을 전제로 AIA를 설계한다. 민간이나 학술연구소에서 의료분야 혹은 공공분야와 같이 특정 분야로 범위로 제한하여 AIA를 개발하는 것(Groves, 2022)과 대조적으로 국가 정부에서 발표하는 AIA의 경우 대부분 공공과 민간에서 모두 사용 가능한 포괄적인 범위의 AIA를 설계를 목표로 한다.

AI 시스템의 수명주기에서 AIA의 개입 시점과 관련하여 독일은 AI 수명주기 전반에 걸쳐 개입해야 설명한다. EU는 생활 전반에 영향을 미친다고 판단되는 AI 시스템이 개발되거나 설계, 적용 시 AIA가 개입해야 한다고 설명한다. 뉴질랜드는 시민 생활에 영향을 미칠 수 있는 알고리즘이 공공서비스에서 사용될 때 AIA를 수행해야 한다고 평가한

다. 미국 NIST는 AI 수명주기의 다양한 단계에서 수행되도록 설계된 반복적이고 지속적인 프로세스이지만 개별 조직의 일정/관심사에 따라 달라질 가능성이 있다고 설명한다. 네덜란드는 AIA는 알고리즘 솔루션 사용에 대한 의사결정 시 필요하며 시스템 사용 전에 개입할 것을 권고한다. 캐나다는 프로젝트 설계 단계 시와 시스템 배포 전 두 번의 평가가 필요하다고 설명한다.

평가수행 방식에 있어 캐나다, 네덜란드는 기본적으로 폐쇄형과 개방형이 혼합된 질문목록을 기본으로 하고 있다. 네덜란드는 정부 조직이 알고리즘 개발, 개발 위임, 구매, 사용을 고려하는 모든 경우에 논의되어야 하는 질의응답 형태를 하고 있다. 캐나다의 경우 질문목록을 제시하는 것은 같다. 그러나 질문목록 작성 후 설문지를 기반으로 시스템의 위험 수준을 계산하는 채점 시스템을 제공한 후, 결과점수에 따른 권장 사항을 제시한다. 독일과 유럽은 AI 시스템의 위험을 분류하고 해당 사항에 맞추어 권고사항을 제시하는 방식을 취하고 있으며 뉴질랜드는 위험 매트릭스를 통한 개방형 자체 평가 방식을 취하고 있다. 미국은 AI 시스템의 설계, 개발, 배포 및 사용 시 위험을 관리, 매핑, 측정, 관리하는 방법을 프로세스별로 제안하는 권고형식을 따르고 있다.

규제방식의 경우 캐나다를 제외한 AIA 사용은 의무가 아닌 자발적 형식을 보이며 대부분 법적 효력 없는 권고사항의 형태를 띠고 있다. 캐나다는 2019년 자동화된 의사결정 지침 재정위원회 훈령을 발표하여 공공기관의 AI 요건을 법규화했다. 해당 훈령은 2020년 4월 이후 캐나다 공공기관에서 사용하는 모든 ADS에 적용되고 있다. ADS 관련 AI 알고리즘을 도입한 공공기관은 영향평가를 시행한다. 사용하는 ADS는 위험의 수준에 따라 분류 및 관리되며 완료한 AIA는 정부 포털에 공개하고 있다. 뉴질랜드는 헌장에 서약한 공공기관을 대상으로 영향평가를 요구하지만, 뉴질랜드의 위험 매트릭스는 특정 질문이 없으며 영향과 위험을 정량화하기 위해 영향평가를 수행하는 실무자에게 큰 자율성을 부여하는 것을 특징으로 한다.

〈표 4〉 AIA 기본구조 분석

| 국가 | 평가시스템 | 평가영역 | 개입시기 | 평가방식 | 규제방식 |
|---|---|---|---|---|---|
| 독일 | 알고리즘 시스템 | 사회 전반 | AI 수명주기 전 단계 | 권고사항 | 자발적 |
| 캐나다 | ADS | 공공기관 | 1)설계 단계<br>2) 배포 전 | (폐쇄형)<br>질문방식 | 의무화 |
| 뉴질랜드 | AI 알고리즘 | 공공서비스 | 배포 후 | 위험 매트릭스 | 의무화 |
| EU | AI 시스템 | 공공부문 | 개발, 설계, 배포 | 권고사항 | 자발적 |
| 미국 | ADS | 사회 전반 | AI 수명주기 전 단계 | 권고사항 | 자발적 |
| 네덜란드 | 알고리즘 시스템 | 공공서비스 | 배포 전 | (폐쇄형/질문형)<br>질문방식 | 자발적 |

## 2. 위험평가 및 영향평가

〈표 6〉은 AIA가 어떻게 AI 시스템이 위험의 수준을 분류하고, 식별하고 있으며, AI 시스템 사용이 개인에 미치는 영향의 성격은 무엇인지 분석하고 있다. 미국을 제외한 5개국은 위험의 수준을 자체적으로 분류하고 있다. 독일은 AI 알고리즘의 위험 수준을 5단계 위험도 피라미드 모델로 제시한다. 해당 모델은 알고리즘 시스템의 중요성에 따라 규제 정도를 조정한다. 중요도가 높은 시스템에는 더 엄격한 기준을 적용하며, 시스템의 중요성은 해당 시스템의 사용으로 발생할 수 있는 피해의 가능성과 심각성을 바탕으로 평가한다. EU는 AI 시스템을 4가지 위험 수준으로 분류하고 있다. 위험별로 구체적인 구성요소는 제시하지 않았지만, 규제 방안을 제시하고 있어 각 분류에 따라 AI 시스템을 다르게 취급할 것을 제안하고 있다. 뉴질랜드의 세 가지 위험등급(낮음, 보통, 높음)으로 분류된다. 기관이 사용하는 AI 시스템의 위험등급이 높을수록 보다 많은 지원을 집중하고 헌장 업무 규칙을 적용하지만, 통상적인 업무 수행 업무 위주로 사용하는 AI 시스템의 경우, 위험등급이 낮으면 대부분의 업무 규칙을 제외할 수 있게 한다. 캐나다는 위험 수준을 4단계로 분류하며 요구사항과 질문을 설정하는 형식이다. 영향평가 결과에 따라 공공기관 AI 시스템의 수준이 결정되면 단계별로 전문가 검토, 공지, 인적 개입, 설명, 검사 모니터링 교육훈련 비상계획, 시스템 구동 승인 의무 등 훈령 상 요건을 차등 적용하도록 하고 있다. FRAIA는 AI 시스템이 기본권에 미치는 위험 정도에 따라 3단계로 위험 단계를 분류하고 있다. 미국의 NIST는 AI 수명주기의 다양한 단계에서 AIA 수행이 가

능하도록 설계된 반복적이고 지속적인 프로세스로 AI의 위험 수준을 분류하지 않는다. 〈표 5〉는 국가별로 AI 위험 수준을 구체적인 내용을 요약하여 정리한 것이다.

〈표 5〉 AI 위험 수준 분류

<table>
<tr><th>국가</th><th colspan="3">AI 위험수준 분류</th></tr>
<tr><td rowspan="3">독일</td><td rowspan="3">5단계</td><td>조치 없음</td><td>레벨1. 해악의 발생 가능성이 전혀 없거나 무시할 수준의 애플리케이션</td></tr>
<tr><td>규제 적용</td><td>레벨2. 해악의 발생 가능성이 약간 있는 애플리케이션<br>레벨3. 해악의 발생 가능성이 정례적이거나 중대한 수준의 애플리케이션<br>레벨4. 해악의 발생 가능성이 심각한 수중의 애플리케이션</td></tr>
<tr><td>사용 금지</td><td>레벨5. 해악의 발생 가능성이 수용 불가능한 수준의 애플리케이션</td></tr>
<tr><td>캐나다</td><td>4단계</td><td colspan="2">레벨1. 의사결정이 (개인 또는 공동체의 권리, 건강, 복리, 경제적 이익, 생태계의 지속가능성)에 거의 영향을 미치지 않을 것으로 보이는 경우<br>레벨2. 의사결정이 (개인 또는 공동체의 권리, 건강, 복리, 경제적 이익, 생태계의 지속가능성)에 중간 정도의 영향을 미칠 것으로 보이는 경우<br>레벨3. 의사결정이 (개인 또는 공동체의 권리, 건강, 복리, 경제적 이익, 생태계의 지속가능성)에 큰 엉향을 미칠 것으로 보이는 경우, (해당 결정은 복구되기 어려울 수 있고 지속적인 영향을 미칠 수 있음<br>레벨 4. 의사결정이 (개인 또는 공동체의 권리, 건강, 복리, 경제적 이익, 생태계의 지속가능성)에 매우 큰 영향을 미칠 것으로 보이는 경우 (해당 결정은 복구가 불가능하고 영구적인 영향을 미칠 수 있음)</td></tr>
<tr><td>뉴질랜드</td><td>3단계</td><td colspan="2">낮음: 알고리즘 헌장이 적용될 수 있음<br>중간; 알고리즘 헌장이 적용됨<br>높음: 알고리즘 헌장이 반드시 적용되어야 함</td></tr>
<tr><td>EU</td><td>4단계</td><td colspan="2">용납할 수 없는 위험<br>고위험<br>제한된 위험(투명성 의무 대상)<br>위험 최소화 또는 위험 없음</td></tr>
<tr><td>네덜란드</td><td>3단계</td><td colspan="2">기본권에 미치는 영향 수준에 따라<br>1단계 심각한 간섭으로 인해 정당화가 필요한 강력한 이유<br>2단계: 중간정도의 간섭으로 주의가 필요<br>3단계: 간섭이 심각하지 않으면 특별한 주의가 필요없음</td></tr>
<tr><td>미국</td><td colspan="3">AI 수명주기 전체</td></tr>
</table>

AI 위험식별에서 데이터의 품질과 관련하여 독일, 캐나다, 미국, 네덜란드는 데이터의 민감도, 적절성, 적시성 여부 등을 질의응답에 포함하고 있다. AI 시스템의 위험평가와 관련하여 투명성과 설명 가능성을 확인하기 위한 AIA 평가 요소는 캐나다와 네덜란드에서만 구체적으로 확인할 수 있다. 캐나다의 자동화된 의사결정에 대한 지침(Directive on Automated Decision-Making 2021)의 제6장 요구사항(Requirements)에서 사용 중인 모든 AI 서비스를 대상으로 한 결정에 대하여 사전공지와 사후 설명을 명시하고 있으며 구성요소에 대한 접근권한, 소스 코드 공개, 전문가 검토 등도 구체적으로 포함하고 있다. 네덜란드의 경우 AI 시스템의 투명성과 설명 가능성에 대하여 개방형 질문을 2B.4(예: 2B.4.4 Transparency and Explainability: 알고리즘의 작동이 질문 B.4.3에서 식별된 대상 그룹에 대해 충분히 이해할 수 있는 방식으로 설명될 수 있습니까?)하고 있다.

AI 알고리즘의 사용이 미치는 영향과 관련하여 뉴질랜드만이 영향이 발생할 가능성과 영향의 심각성 정도를 정성적으로 평가하고 있다. 뉴질랜드의 위험평가 매트릭스는 AI 알고리즘 사용이 사람에게 미치는 영향을 발생빈도(통상 있음, 때때로 있음, 거의 없음)와 심각성(낮음, 중간, 높음)으로 평가하며 이를 토대로 헌장의 적용 여부를 결정한다. 영향의 성질과 관련하여 독일, 네덜란드, EU는 법적 권리, 신체적/정신적 안녕을 중점적으로 채택하고 있다. 여기에는 개인정보에 관한 권리, 생명 및 신체적 완전성에 대한 기본권, 차별 금지 등이 포함되어 있으며 비물질적/정신적 피해 또는 금전적으로 계산하기 어려운 효용 손실을 포함한 잠재적 피해의 수준 등을 포함하고 있다. 뉴질랜드는 공공서비스 중 복지 분야에서 사용하는 AI 알고리즘에 대한 영향평가를 중심으로 수행하고 있어 신체적/정신적 안녕에 영향을 미치는 부분을 채택하고 있다. 미국 NIST는 영향의 범주를 긍정적인 영향과 부정적인 영향으로만 구분하여 설명하고 있다. 캐나다의 ADS의 위험성을 수준별로 나누어 관리하고 있는데 수준을 구분하는 세부 사항의 조건에 ADS의 사용이 개인 또는 공동체의 권리, 개인 또는 공동체의 건강 또는 복리와 함께 개인 전체 또는 공동체의 경제적 이익에 영향을 미치는지 확인하고 있다.

〈표 6〉 AIA의 위험평가 및 영향평가 분석

| 국가 | 위험 평가 | | | 영향 평가 | | |
|---|---|---|---|---|---|---|
| | 위험수준 | 위험식별 | | | | |
| | | 데이터 | 시스템 | 빈도 | 심각성 | 영향의 성질 |
| 독일 | 5단계 | ○ | - | - | - | 법적 권리에 영향<br>신체적/정신적 안녕에 영향 |
| 캐나다 | 4단계 | ○ | ○ | - | - | 법적 권리에 영향<br>신체적/정신적 안녕에 영향<br>경제적 안정에 영향 |
| 뉴질랜드 | 3단계 | - | - | ○ | ○ | 신체적/정신적 안녕에 영향 |
| EU | 4단계 | - | - | - | - | 법적 권리에 영향 |
| 미국 | AI 수명주기 전체 | ○ | - | - | - | 긍정적인 영향/부정적인 영향 |
| 네덜란드 | 3단계 | ○ | ○ | - | - | 법적 권리에 영향<br>신체적/정신적 안녕에 영향 |

## 3. 영향평가의 위험 완화조치 분석

〈표 7〉은 AIA에서 AI 시스템 사용 시 발생 가능한 위험의 완화조치와 관련한 주요 내용을 분석하고 있다. 위험 완화조치와 관련하여 편향 테스트, 데이터 관리를 통한 편향 완화조치, 외부 이해관계자의 참여 및 검토, 주기적 혹은 지속적 모니터링에 대하여 6개국 모두 완화조치로 포함하고 있다. 최종결정 단계에서 사람의 필수적인 개입은 뉴질랜드를 제외한 5개국 모두 완화조치로 포함하고 있다. 외부 이해관계자 참여 및 검토는 6개국 모두 완화조치에 포함하였지만, 대중의 공개 여부, 지역사회의 참여, 이의제기 채널 구성과 같은 구체적인 내용에 대해서는 대부분 포함하지 않았다. 독일만이 시정 또는 이의제기를 위한 채널 설정을 완화조치에 포함하고 있으며 EU와 뉴질랜드는 AI 시스템의 알고리즘을 대중의 공개하는 것을 위험 완화조치에 포함하고 있다. 마지막으로 사용금지와 관련하여 몇몇 국가는 위험분류 단계에 따라 가장 높은 위험 수준의 AI 시스템의 사용을 금지하고 있다. 독일은 해악의 발생 가능성이 수용 불가능한 수준의 애플리케이션(5단계)의 경우 알고리즘 시스템의 전체 또는 부분적으로 금지하고 있다. EU는 잠재의식 조작, 아동·장애인 착취, 사회적 평점 시스템, 실시간 원격 생체정보기반 식별과 관련한

시스템 등을 용납할 수 없는 위험(Unacceptable Risk)의 단계에 포함하고 있다. 해당 단계의 AI 시스템 규제 방안과 관련하여 잠재의식 조작, 아동·장애인 착취, 공적 범용 사회적 평점 시스템의 경우 출시·서비스·활용을 모두 금지하고 있다. 실시간 원격 생체정보기반 식별의 경우 법을 집행하기 위한 목적으로 공공장소에서 실시간으로 활용은 금지하지만, 기타 활용이나 출시·서비스는 기록보존 및 인적 감시와 관련된 추가적인 요건을 부과하고 있다.

〈표 7〉 AIA 위험 완화 조치

| 국가 | 편향완화 | 인간개입 | 외부자 개입 | | | 모니터링 | 사용금지 |
|---|---|---|---|---|---|---|---|
| | | | 외부자 참여 | 대중공개, 지역사회협력 | 이의제기 채널설정 | | |
| 독일 | ○ | ○ | ○ | - | ○ | ○ | ○ |
| 캐나다 | ○ | ○ | ○ | - | - | ○ | - |
| 뉴질랜드 | ○ | - | ○ | ○ | ○ | ○ | - |
| EU | ○ | ○ | ○ | ○ | - | ○ | ○ |
| 미국 | ○ | ○ | ○ | - | - | ○ | - |
| 네덜란드 | ○ | ○ | ○ | - | - | ○ | - |

## 4. 분석 함의

본 연구는 연구 질문과 관련하여 다음의 경향성과 특징을 확인했다. 첫 번째 연구 질문은 다음과 같다. 공공분야에서 발표한 AIA의 공통점, 차이점, 주요 특징은 무엇인가?

첫째, 기존 영향평가와의 차이이다. AIA는 기존 영향평가와 다른 기술적·산업적 특수성을 인정해야 한다는 인식이 존재한다. AI 시스템은 일반적으로 전통적인 영향평가에서 가정하는 것처럼 정적이지 않으며 끊임없이 새로운 데이터를 추가하고, 모델을 학습하고, 구체화하는 동적인 특성이 있다(Calvo 외, 2020). 이는 AIA를 설계하고 구현하는데도 적용된다. 즉, 기존 다른 분야의 영향평가들이 규모가 방대하고 오랜 시간이 걸리는 것과 대조적으로 공공기관에서 구현한 AIA는 상대적으로 짧은 시간 내에 답할 수 있는 포괄적 형태의 간단한 AIA 유형이 다수를 차지한다.

둘째, AI의 개념과 정의에 관한 것이다. AI에 대한 정의는 광범위하게 논의되고 있지

만, 보편적으로 수용되는 명확한 정의가 없으며(Elsevier 2018), AI의 위험을 평가하는 AIA 역시 정확한 범위를 정의하기 어렵다. AIA를 개발하는 국가 정부 역시 AI 시스템을 포괄적이고 광범위하게 정의한다. 이는 모든 유형의 AI 시스템 위험과 영향에 대한 평가를 고려해야 한다는 것을 의미한다. 즉, 평가의 대상이나 영역이 다양해질 뿐만 아니라, 사용 주체 및 유용성의 허용범위에 따라서 결과가 크게 달라질 수 있다는 것을 의미한다. 이러한 광범위하고 불분명한 용어 정의는 새로운 모호성을 유발하며 AIA를 통해 해결하고자 하는 다양한 AI 시스템의 문제를 해결하지 못할 우려가 있다.

셋째, 국가 정부와 공공기관에서 발표한 AIA는 구조, 내용, 구현에 있어 많은 합의를 이루지 못하고 있다. 그러나 위험 완화조치와 관련하여 일부 의견이 수렴하고 있는 것을 확인하였다. 대부분의 AIA는 위험의 수준을 분류하고 있다. 또한, 편향완화, 인간의 개입 및 감독, 모니터링 등 위험 완화와 관련한 많은 세부 내용에서 의견일치를 보인다. 그러나 외부 이해관계자 검토와 관련하여 대중의 공개 여부, 지역사회의 참여, 이의제기 채널 구성과 같은 구체적인 내용에서는 대체로 의견이 일치하지 않았다. 또한, 데이터와 시스템의 위험식별과 관련해서도 크게 일치하지 않았다. 이러한 불일치의 대부분 AIA의 투명성, 설명 가능성, 해석 가능성과 관련한 사항으로 국제기구, 산업기관, 시민단체와의 적극적인 협조와 합의가 필수적인 영역이다. AIA와 관련하여 보편적으로 수용되는 모델이나 공통 규제 프레임워크를 개발하기 위해서는 앞에서 언급한 해당 영역에서 다양한 이해관계자들과의 추가적인 논의가 필요할 것으로 보인다.

넷째, AIA에서 제시하는 완화조치가 위험분류 수준과 연결되는 방식이 국가마다 상이하다. 예를 들어, 캐나다 정부 지침과 뉴질랜드의 알고리즘 헌장의 경우 위험 수준이 높을수록 더 많은 완화조치가 필요하며 조치의 요구사항도 엄격해진다. 구체적으로 높은 위험 수준의 AI 시스템일수록 시스템에 필요한 검토 수가 증가한다. EU 국가들의 경우, 위험이 낮거나 최소한으로 분류된 AI 시스템의 경우 특별한 조치가 필요하지 않다. 그러나 높은 수준의 위험을 초래하는 시스템이라고 판단하는 경우, 혹은 현재의 AI 기술의 발전 수준과 제도 수준에서 정부 조직이 해당 기술에 적절한 조처를 하는 것이 어렵다고 판단될 경우는 시스템 사용을 금지한다.

본 연구의 두 번째 연구 질문은 다음과 같다. 학술적으로 제안하는 AIA와 정책으로 구현된 AIA는 어떠한 차이가 있는가?

첫째, 개입 시기에 대한 차이이다. 현재 정책으로 구현한 AIA는 AI 수명주기의 여러 지점에서 수행된다. 그러나 선행연구는 AIA의 조기 개입을 제안한다. AI를 외부에서 구매하는 경우 배포 전에 AIA를 구현할 것을 권고하며, AI를 내부에서 자체적으로 설계하고 개발하는 경우 프로젝트 초기에 AIA를 수행할 것을 권장한다(Reisman, 2018; Selbst, 2021; PwC, 2021). AI 시스템은 다양한 설계 요소가 미묘한 방식으로 상호작용하여 결정에 이르며, 배경 근거를 포괄적으로 문서화하지 않으면 오류에 대한 인과관계 설명을 정확히 찾아내기 어렵다. 그러나 AIA가 조기 개입 할 경우, 문서를 통해 취해질 수 있는 대체 의사결정 경로를 추적할 수 있다. 차별, 설명, 안전 등의 문제를 해결하기 위해 AI 시스템 결과를 이해하려면 시스템 개발 중에 내려진 주관적인 결정에 대한 정보 확인이 필요하다. 이는 결정을 정당화하는 데 도움이 되며, 문제가 발생할 때 시스템 작동을 해결하는 데 중요한 역할을 한다. 이처럼 선행연구가 주장하는 AIA의 조기 개입은 복잡한 AI 시스템이 개발되어 완성된 이후에는 결과를 설명하는 것이 어렵거나 불가능함을 전제하고 있다(PwC, 2021). 따라서 현재 정책으로 구현되었거나 제안된 AIA는 연구에서 제시하는 AIA보다 영향평가 수행의 효과가 상대적으로 적을 수 있음을 추측할 수 있으며, 현재의 제도적 환경에서 강력한 AIA 사전영향평가 수행이 어렵다는 것을 간접적으로 가늠할 수 있다. 이는 효과적인 AIA 수행을 위해서는 영향평가 자체의 효과성은 물론 제도적 변화가 동반되어야 함을 시사한다.

둘째, 규제방식에 관한 차이이다. 정책으로 구현하거나 제안하는 AIA는 자율규제적 성격이 강하다. 기존 연구에서는 강력한 법적 조치와 책임을 강조하지만(Selbst 2021), 현재까지 국가 정부와 공공기관에서 발표한 AIA는 대부분 책임에 대한 강제력을 포함하지 않고 있으며 자발적 참여를 통한 권고사항 수준에 그치고 있다. 검토된 대부분의 AIA는 공공 참여를 위한 명확한 프로세스, 결과에 대한 투명성, 설명 가능성, 해석 가능성, 대중 혹은 외부 접근을 보장하는 명확한 절차, AI 시스템 개발 및 배포와 관련하여 직접적이고 명확한 책임 라인을 설정하지 않는다. 학술연구에서 AIA는 AI 시스템의 구현과정에서 이루어진 특정 선택을 설명하거나 정당화할 수 있는 공식적인 답변 요구가 가능한 것을 전제한다. 따라서 이러한 요구를 공식적으로 수렴하는 대표 기관 혹은 위원회의 존재를 가정한다(Reisaman, 2018). 이들은 구현된 시스템의 목적과 의도, 평가 결과 간의 차이 혹은 식별된 문제점에 대해 강제할 수 있는 권한을 가진 행위자로 가정한다. 또 다

른 경우에는 외부 기관에 의해 시행되는 규제적 대응을 통해 책임을 질 수 있다. 그러나 정책으로 구현하거나 제안한 대부분의 AIA는 확인된 위험에 대해 조치를 할 수 있는 법적 강제력이나 정치적 권한이 없다. 또한, 공공기관에 책임을 묻기 위한 명확한 조처를 설정하지 않는다. 대부분 AIA는 기관의 자체 평가를 목적으로 하는 재량의 영역에 속하며 공식적으로 공개할 것을 요구하지 않는다. 본 연구에서 살펴본 국가에서 발표한 AIA는 대체로 자발적 참여를 권고하는 자율규제 성격을 보이며, 외부 행위자나 책임 집행을 위한 포럼에 대한 언급 없이 공공기관 내부적으로 의사결정 지침을 제공하는 데 사용하고 있다.

# V. 결론

AIA는 공공기관에서 활용하는 새로운 정책 메커니즘으로 AI 시스템 사용으로 발생할 수 있는 잠재적 피해나 위험을 식별하고 분류하고 대응하는 조치 도구이다(Ada Lovelace Institute 외, 2021). 본 연구는 AI 시스템의 위험 완화를 위한 정책접근 방식으로 2023년 현재까지 공공분야에서 발표한 AIA에 대한 탐색적 사례분석 수행을 목표로 한다. 이를 위해 본 연구는 선행연구에서 제시한 AIA 설계의 핵심 고려사항을 토대로 연구분석 틀을 구성한다. 이후 분석 틀을 토대로 현재까지 국가 정부 혹은 공공기관에서 정책으로 구현하거나 제안한 AIA 사례를 채택하여 비교 분석하였다.

다양한 AIA 접근방식이 선행연구를 통해 이론적으로 제시되었음에도 구체적이고 운영 가능한 AIA 개발 사례는 여전히 부족하다. 또한, 공공분야에서 AIA의 도입은 초기 단계로 일관된 방법론이 없고, 사용을 요구하는 법적 기반이 없으며, 이러한 연습을 수행할 권한이 있는 평가자에 대한 구체적인 시장이 없다. 책임성, 투명성, 공정성과 같은 AI 시스템의 개발 및 사용에 대한 원칙의 중요성에 대한 합의가 높아지고 있지만 이러한 용어에 대한 우선순위와 해석은 조직마다 다르다. 또한, AI 시스템이 영향을 미치는 정도, 영향의 심각성, 위험의 식별, 투명성과 설명 가능성의 구현 범위 여부도 다른 견해를 가지고 있다.

그러나 AIA는 엄격한 법률 규정이 없더라고 책임감 있는 AI 개발을 장려하는 중요한

접근방법으로 윤리적, 사회적 문제를 포함하여 AI 사용 시 발생 가능한 위험을 완화하기 위한 실용적인 접근방식을 제공한다는 데 큰 장점이 있다(Stix, 2021). 또한, AIA는 현재의 피해를 완화하고 미래에 더 효과적이고 구체적인 규제를 개발하기 위한 수단이 될 수 있다. AIA를 사용하면 AI 시스템이 개발되어 사회에 배포될 때 AI 시스템의 위험을 측정하고 잠재적 피해를 완화하기 위한 적절한 조치를 할 수 있으며 이 과정에서 충분한 정보를 얻을 수 있다. AIA 사용 규제환경에서 만들어지는 관련 정보와 문서작업은 AI 사용과정을 투명하고 책임감 있게 운영하는데 긍정적인 작용을 한다.

해당 연구는 실질적이고 구체적인 지침을 제공하기 위해 공공기관에서 제안한 AIA를 비교분석 했다. AIA 개발과 관련하여 현재까지의 정책 동향을 살펴봄으로써 AIA 개발에 대한 전반적인 이해를 도울 수 있는 포괄적인 자료를 제공하고, 추후 정책개발에 정보를 제공할 수 있다는 점에서 본 연구의 의의가 있다. 그러나 해당 연구는 다음과 같은 한계를 갖는다. 첫째, 사례의 대표성 부족이다. 다양한 AIA 방식이 발표되었음에도 구체적이고 운영 가능한 AIA는 여전히 부족하다. AIA에 대한 다양한 접근방식이 제안되었지만, 대부분 광범위한 합의를 얻지 못했다. 또한, 여전히 구체적인 수행 사례가 부족하며 실질적인 정책 결정 커뮤니티에 참여하지 못하고 있다. 공공의 영역에서 제안했다는 공통점을 제외하고는, AIA의 형식, 내용, 접근방식의 차이가 큰 해당 사례들을 동등하게 비교한 것이 연구의 한계라고 할 수 있다. 둘째, 위험분석 식별에서 데이터와 시스템 지표의 구체적인 분석을 연구내용에 포함하지 못해 기술분석에 있어 탐색적 수준에 그쳤다.

해당 연구분석의 결론을 통해 연구자가 얻은 추가적인 연구 질문은 다음과 같다. 첫째, 조직은 내부 자체 평가를 위해 영향평가 도구를 사용하는 것을 넘어 투명성, 설명 가능성, 해석 가능성을 높이기 위해 어느 정도까지 AI 시스템의 알고리즘을 외부자 또는 대중에게 공개할 수 있는가? 둘째, 정부가 규제할 수 있는 방식으로 AI 시스템의 위험과 사회적 영향을 어떻게 측정할 수 있는가? 후속 연구로는 해당 질문을 바탕으로 연구를 진행하고 의미 있는 정책 방향을 제시하고자 한다.

## 참고문헌

김근혜, 박규동. (2022). AI 영향평가에 관한 국외 연구 동향 분석. 차세대융합기술학회논문지. 6(3): 615-623.

권은정. (2023a). 인공지능 서비스 영향평가의 체계와 방법론 -「지능정보화 기본법」의 사회적 영향평가 제도를 중심으로-. 경제규제와 법, 16(1): 29-50.

권은정. (2023b). 자동적 처분에 대한 통제 수단으로서의 알고리즘 영향평가. 국가법연구, 19(2): 121-148.

김법연. (2023). 공공분야 인공지능서비스의 영향평가제도 도입에 관한 연구. 정보법학, 27(2): 171-219.

김진열. (2020). 정책 '영향평가제도'와 「소비자영향평가」에 대한 소고. 소비자정책동향, 109: 1-25.

노화준. (1986). 프로그램 이론 형성으로서의 정책영향평가. 행정논총, 24(1): 47-64.

유순덕. (2023). 인공지능 서비스 영향성 평가를 위한 분석 기준 연구. 한국인터넷방송통신학회논문지, 23(1): 7-13.

Ada Lovelace Institute. (2020). Examing the Black Box. (https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/)

Ada Lovelace Institute, AI NOW Institute, Open Government Partnership. (2021). Algorithmic Accountability for the Public Sector: Learning from the First Wave of Policy Implementation. (https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/)

Calvo RA, Peters D, Cave S (2020) Advancing impact assessment for intelligent systems. Nat Mach Intell 2:89-91

Chae. Y. (2020). US AI regulation guide: legislative overview and practical considerations. The Journal of Robotics, Artificial Intelligence & Law, 3.

Data Ethics Commission. (2019). Opinion of the Data Ethics Commision. (https://www.bmj.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2).

Elsevier. (2018). Artificial intelligence: how knowledge is created, transferred, and used-trends in China, Europe, and the United States. Elsevier, Amsterdam.

European Commission. (2017). Completing the Better Regulations Agenda: Better Solutions for better Results. (https://commission.europa.eu/system/files/2017-10/completing-the-better-regulation-agenda-better-solutions-for-better-results_en.pdf)

European Commission. (2021). Regulatory Framework Proposal on Artificial Intelligence. (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai)

Ezeani, G., Koene, A., Kumar, R., Santiago, N., & Wright, D. (2021). A Survey of Artificial Intelligence Risk Assessment Methodologies. Technical Report, Ernst & Young LLP.

Government of Canada. (2019). Directive on Automated Decision-Making. (https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592).

Government of the Netherlands. (2022). Fundamental Rights and Algorithms Impact Assessment (FRAIA). (https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms)

Groves, L. (2022). Algorithmic impact assessment: a case study in health care. Technical report, AdaLovelace Institute. (https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/)

IFOW(Institute for the Future of Work). (2021). Policy Briefing Algorithmic Impact Assessments. (https://www.ifow.org/publications/policy-briefing-building-a-systematic-framework-of-accountability-for-algorithmic-decision-making)

Kaminski, M. & Malgieri, G. (2019). Algorithmic impact assessments under the

GDPR: producing multi-layered explanations. International Data Privacy Law, 19-28.

Kazim E. & A. Koshiyama. (2021). The interrelation between data and AI ethics in the context of impact assessments. AI and Ethics, 1(3), 219-225. DOI: 10.1007/s43681-020-00029-w

Kazim, E., D. Denny & A. Koshiyama. (2021). AI auditing and impact assessment: according to the UK information commissioner's office. AI and Ethics, 1(3), 301-310.

Long, S. (2022). Explainer: Impact Assessments for Artificial Intelligence. (https://bipartisanpolicy.org/blog/impact-assessments-for-ai/)

Lovelace, A., & DataKind, U. K. (2020). Examining the black box: Tools for assessing algorithmic systems. Technical report, AdaLovelace Institute.(https://ico.org.uk/media/about-theico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.)

Lo P, S. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanities and Social Sciences Communications. 7(1): 1-7.

Mantelero, A. (2018). AI and Big Data: a blueprint for a human rights, social and ethical impact assessment. Comput Law Secur Rev 34:754-772. https://doi.org/10.1016/j.clsr.2018.05.017.

Marquenie, T., & Quezada-Tavárez, K. (2022). Data Protection Impact Assessments in Law Enforcement: Identifying and Mitigating Risks in Algorithmic Policing. Security Technologies and Social Implications, 32-60.

Nahmias, Y. & M. Perel. (2021). The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations. Harv. J. on Legis., 58, 145.

New Zealand Government. (2020). Algorithm Chater for Aotearoa New Zealand. (https://data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-2020_Final-English-1.pdf).

NIST(National Institute of Standards and Technology). (2023). Roadmap for the NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0). (https://www.nist.gov/itl/ai-risk-management-framework/roadmap-nist-artificial-intelligence-risk-management-framework-ai)

OECD(Organisation for Economic Co-operation and Development). (2019). OECD Principles on AI. (www.oecd.org/going-digital/ai/principles).

OECD. (2020). Regulatory Impact Assessment. (https://www.oecd.org/gov/regulatory-policy/regulatory-impact-assessment-7a9638cb-en.htm).

OECD. (2021). Tools for Trustworthy AI: A Framework to Compare Implementation Tools for Trustworthy AI Systems. (https://www.oecd.org/science/tools-for-trustworthy-ai-008232ec-en.htm).

OECD. (2023). The state of implementation of the OECD AI Principles four years on. (https://www.oecd.org/publications/the-state-of-implementation-of-the-oecd-ai-principles-four-years-on-835641c9-en.htm)

PwC(PricewaterhouseCoopers International Limited). (2021). Algorithmic impact assessments: What are they and why do you need them?. (https://www.pwc.com/us/en/tech-effect/ai-analytics/algorithmic-impact-assessments.html)

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic Impact Assessments: A Practical Framework for Public Agency. AI Now. (https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2)

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 44(1.2), 206-226.

Selbst, A. D. (2021). An Institutional View of Algorithmic Impact. Harvard Journal of Law & Technology, 35(1): 117-191.

Shah, H. (2018). Algorithmic accountability. Philosophical Transactions of the

Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2128), 20170362.

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., ... & Wright, D. (2023). A Systematic review of artificial intelligence impact assessments. Artificial Intelligence Review, 56: 12799-12831.

Stix, C. (2021). Actionable principles for artificial intelligence policy: three pathways. Science and engineering ethics, 27(1), 15.

Thierer. A. (2023). NEPA for Al? The Problem with Mandating Algorithmic Audits & Impact Assessments. Medium. (https://medium.com/@AdamThierer/nepa-for-al-the-problem-with-mandating-algorithmic-audits-impact-assessments-8eeceaea64ad)

Wiener, N. (1988). The human use of human beings: cybernetics and society. Da Capo Press, New York, N.Y, new edition.

Yam, J. & J. Skorburg. (2021). From human resources to human rights: Impact assessments for hiring algorithms. Ethics and Information Technology, 23(4), 611-623.

Yeung, L. A. (2021). Guidance for the development of AI risk and impact assessments. Center for Long-Term Cybersecurity, University of California, Berkeley.

# A Study on Policy Approaches to Mitigate Risks in AI System : Focusing on AI Impact Assessment

Geun Hye Kim, Kyu Dong Park

This study aims to conduct an exploratory case analysis of AIA(AI Impact Assessment) published in the public sector as a policy approach to mitigating the risk of AI systems. This study constructs a research analysis framework based on the key points of AIA design presented in previous studies. Based on the subsequent analysis framework, six AIA cases implemented or proposed as policies by national governments or public institutions to date have been adopted and analyzed. As a result of the analysis, there was a perception that AIA should recognize technical and industrial specificity different from the existing impact assessment, and the AI system was comprehensively and broadly defined. Although some opinions were gathered regarding risk mitigation measures, the way the mitigation measures presented in the AI impact assessment were linked to the level of risk classification differed from country to country. In addition, it was carried out regularly at various points in the AI life cycle and had a strong self-regulatory character. This study contributes to a deeper understanding of AIA development by presenting comprehensive data on policy and research trends, thereby informing future policy development in this field.

Keyword: Artificial Intelligence, AI Impact Assessment, Algorithmic Assessment, Automated Decision System