기획논문 「정부학연구」 제30권 제2호(2024): 61~84
DOI https://doi.org/10.19067/jgs.2024.30.2.61

원고투고일: 2024.07.15.
게재확정일: 2024.08.09.

누구를 위한 디지털 전환인가? 자동화된 복지행정의 위험성*

홍승헌**
황하***

인공지능 등 신기술에 기반한 행정서비스의 급속한 도입은 기회와 위기를 동시에 가져오고 있다. 한편으로는 과도한 자원과 인력이 소모되던 방대한 데이터의 수집과 처리, 사각지대의 발굴, 보다 과학적인 의사결정 보조, 고객친화적인 서비스 경험 등을 가능케하면서 정책 및 행정서비스 과정을 보다 효율적·효과적으로 만들고 있는 반면, 수집한 개인정보의 처리, 알고리즘 편중(bias)의 문제, 사이버보안과 같은 문제들을 양산하고 있다. 본 논문은 인공지능 기반 행정서비스가 인간에 미칠 수 있는 위험성에 주목하여, 이를 해결하기 위해 인공지능을 설계하고 도입하는 과정에서 어떠한 인간개입이 필요한지를 논한다. 자동화된 복지행정서비스가 사회적 약자인 복지수급자들에게 악영향을 끼친 대표적 사례인 호주의 로보데트 스캔들에 초점을 맞춰서, 인공지능 기반 행정서비스에 내재된 다양한 위험성(risk)에 대해 논의한 후, 이에 대처하기 위하여 인공지능 시스템 설계시 구비해야 할 조건으로 인간의 3중 개입(triple-loop human intervention)을 제시한다.

주제어: 디지털 전환, 복지행정, 인공지능, 로보데트, 인간개입

* 본 논문의 초고는 2024년 6월 고려대 비교거버넌스연구소가 개최한 "디지털 전환의 거버넌스: 돌봄과 복지의 미래" 세미나에서 발표되었습니다. 소중한 비평을 해주신 참석자 선생님들, 본 논문을 심사하신 세 분의 익명의 심사자 선생님, 그리고 정부학연구의 편집자님께 깊은 감사를 드립니다. 이 논문은 2021년 대한민국 교육부와 한국연구재단의 일반공동연구지원사업의 지원을 받아 수행된 연구임을 밝힙니다(NRF-2021S1A5A2A03064391).
** 제1저자: Australian National University, Regulation and Global Governance, 한국행정연구원 규제정책연구실장(연구위원), 규제, 거버넌스, 인공지능 기반 행정서비스, 신공화주의 제도론(E-mail: seunghun.hong@kipa.re.kr)
*** 교신저자: SUNY Buffalo, Urban and Regional Planning, 한국행정연구원 규제정책연구실 연구위원, 재난안전, 기후변화, 인공지능 기반 레그테크(E-mail: ghkdgk0309@gmail.com)

Ⅰ. 서론

전세계적으로 인공지능 기반 행정서비스가 급속도로 늘어나고 있다. 정부분야에서 인공지능 등의 신기술 적용을 총칭하는 거브테크(GovTech) 뿐만 아니라, 신기술을 활용하여 규제준수를 보다 효과적으로 하려는 레그테크(RegTech) 솔루션이 의료, 환경, 고용, 납세, 개인정보, 제품안전 등 다양한 분야에서 도입되고 있다(김이인정 외, 2022; 이종한 외, 2023). 인공지능 등 신기술에 기반한 행정서비스의 급속한 도입은 기회와 위기를 동시에 가져오고 있다. 한편으로는 과도한 자원과 인력이 소모되던 방대한 데이터의 수집과 처리, 기존에 파악하기 어렵던 사각지대의 발굴, 보다 과학적인 의사결정 보조, 고객친화적인 서비스 경험 등을 가능케하면서 정책 및 행정서비스 과정을 보다 효율적·효과적으로 만들고 있는 반면, 수집한 개인정보의 처리, 알고리즘 편중(bias)의 문제, 사이버보안과 같은 문제들을 양산하고 있다.

국내에서 인공지능 기반 행정서비스에 관한 연구는 역사가 그리 길지 않지만 매우 다각적으로 전개되어 왔다. 초기의 연구들은 행정서비스 디지털 전환의 가능성과 기대감에 주로 초점을 맞추고 있다. 인공지능에 기반한 챗봇이 사용자에게 맞춤형 정보를 제공해줌으로써 정부가 국민 개개인과의 소통이 증진될 수 있는 가능성에 주목한 연구(박동안, 2017; 윤상오, 2018; 송진순, 2022), 인공지능 기반 행정서비스가 창출해낼 것으로 기대되는 공공가치에는 어떤 것들이 있는지에 관한 연구(이제복 & 최상옥, 2018; 김길수, 2019) 등이 대표적이다. 이후의 연구들은 인공지능이라는 새로운 기술의 보편적 도입이 가져올 잠재적 갈등요소와 위험성에 보다 주목하여 전개되어 왔다. 인공지능 기반으로 행정서비스가 제공되었을 때의 국민수용성에 주목하거나(한명성, 2021; 장창기 & 성욱준, 2022), 인공지능 기반 행정서비스나 자동화된 행정결정이 지닌 위험성에 대해 논의하는 연구(김대인, 2020; 김휘식, 2021; 안준모, 2021; 조인성, 2022; 김한솔 & 김병조, 2023; 김법연, 2023; 유순덕, 2023) 등이 대표적이다.

본 논문은 인공지능 기반 행정서비스가 인간에 미칠 수 있는 위험성에 주목하는 것을 넘어서, 이를 해결하기 위해 인공지능을 설계하고 도입하는 과정에서 어떠한 인간 개입이 필요한지를 논한다는 점에서 기존의 연구들과 차별성을 지닌다. 특히 본 논문이 주목하는 것은 행정서비스가 지닌 독점성으로 인하여, 행정서비스의 디지털 전환이 가져올 수 있는 해악으로부터 그 결과에 영향을 받는 대상자들이 취약해지기 쉽다는 점이다. 민간이 제공하는 서비스의 경우 서비스에 불만이 있거나 더 좋은 서비스를 발견할 경우 사용자가 해당 서비스를 사용하지 않을(opt-out) 선택을 할 수 있으나, 정부가 공급을 독점하는 행정서비스의 경우 사용자의 탈출 옵션이 매우 제한적이거나,

있더라도 많은 경우 선택하는 것이 사실상 불가능할 수밖에 없다.[1] 따라서 이러한 독점 서비스의 경우, 서비스 수급자들에게 영향을 미치는 정책적·행정적 의사결정을 자동화에 의존하는 것에 매우 신중할 필요가 있다.

인공지능 알고리즘이 내놓는 판단의 정확도가 항상 100%일 수 없기에 우리는 더더욱 신중을 기할 필요가 있다. 알고리즘의 정확도가 99%에 달하더라도 백만 건 중 1만 건 가량의 오류가 발생할 수 있고, 한국의 인구를 고려할 때 이는 곧 수만에서 수십만에 달하는 행정서비스 대상자들에 대하여 잘못된 판단이 내려질 수 있다는 것을 의미한다. 따라서 인공지능 기반 행정서비스를 설계하고 도입하는 데 있어 이러한 오류가 발생할 수 있는 가능성을 재차 점검하고 만일의 경우 발생할 수 있는 피해로부터의 구제절차를 다각적으로 만들어 놓을 필요가 있다. 본 연구는 행정서비스 대상자의 대부분이 사회적 약자일 수 있는 복지행정의 사례를 들어 그 위험성을 논의하고 이를 해결할 수 있는 정책적 대안을 제시하고자 한다.

본 논문에서 주목한 사례는 자동화된 복지행정서비스가 인간, 특히 사회적 약자인 복지수급자들에게 악영향을 끼친 대표적 사례인 호주의 로보데트 스캔들(Robodebt Scandal)이다. 이 사례는 개개인에 대한 복지수당 초과지급분의 계산이 자동화되어 이뤄지고, 이를 기반으로 수급자 개개인에 대한 초과지급분 반환결정 역시 자동화에 의존하면서 발생한 문제들이 종합적으로 드러난 경우다. 호주 정부가 2016년 도입한 로보데트 시스템(Robodebt Scheme)은 고용주가 국세청에 신고한 소득자료와 복지수당 수령을 위해 개인이 자진신고한 소득의 내용이 일치하는 지를 확인한 후, 과다 지급된 복지수당이 있을 시에는 이를 환수하는 조치를 통보하는 자동화된 시스템이었다. 복지수당이 초과지급된 대상을 특정하고 이들에게 부채고지서를 발급하는 데 있어 데이터 매칭 알고리즘이 핵심적 역할을 담당하였다. 그러나, 이러한 과정에서 문제를 확인하고 시스템의 정확성을 높이려는 담당부처의 노력이 부재하였기 때문에 결과적으로 시스템이 내놓은 부채고지서의 부정확성을 입증하는 책임은 사회적 약자인 복지수급자들에게 전가되는 결과를 낳았다(Braithwaite, 2020: 243). 즉, 문제해결의 효율성과 효과성을 제고하기 위하여 자동화된 시스템을 도입하였으나, 이에 과의존함으로써 수많은 사회적 약자들에게 해악(peril)을 끼친 대표적 사례라고 할 수 있다.

본 논문에서는 이러한 복지행정서비스가 가져온 해악에 대한 분석에 기반하여 보다

1) 물론 상업적 인공지능 서비스의 경우에도 개인의 탈출 옵션이 현실화되기가 어렵다는 점은 존재하며(Gotterbarn, 2010), 정부가 제공하는 공공서비스가 다양한 방식으로 다변화되고 있는 점 또한 인정할 필요가 있다(박석희 & 조강주, 2016). 그럼에도 불구하고, 복지수급자의 선정과 복지수당 지급과 같은 일부 행정서비스는 여전히 국가가 공급을 독점하고 있다.

보편적으로 인공지능 기반 행정서비스가 가질 수 있는 다양한 위험성(risk)에 대해 논의한 후, 이에 대처하기 위하여 인공지능 시스템 설계시 구비해야 할 조건에 대하여 제시하고자 한다. 본 논문의 구성은 다음과 같다. 2절은 호주 정부의 복지서비스를 소개하고 로보데트 시스템의 도입배경을 소개한다. 3절에서는 자동화된 복지행정이 지니는 위험성에 대해서 논의하며, 이어 4절에서는 이러한 위험성 대응을 위해 인공지능 서비스가 가져야 할 조건에 대해서 논의한다. 5절에서는 논문의 시사점과 정책적 대안, 논문의 한계에 대해 서술하면서 논문을 마무리할 것이다.

Ⅱ. 호주의 로보데트 스캔들

지난 10년 내 호주에서 가장 큰 사회적 문제 중 하나로 대두되었던 소위 로보데트 스캔들(Robodebt Scandal)은 자동화된 의사결정에 의존하는 것이 인간, 특히 사회적 약자들에게 얼마나 가혹한 영향을 줄 수 있는지를 잘 보여주는 사례라고 할 수 있다. 2016년 7월 본격적으로 도입된 로보데트 시스템(Robodebt Scheme)은 정부가 지급하는 복지수당, 복지급여 등이 과다 지급되었는지 등을 판별하고, 과다수급자들에게 부채환수고지서를 발송하는 자동화된 서비스다. 복지수당수급자들이 복지부에 자진신고한 소득정보를 국세청이 파악한 납세/소득 정보와 연동시켜서 실질 소득에 맞게 과다지급된 복지급여 등을 환수하는 일종의 자동부채환수시스템이라고 할 수 있다. 기존에 담당자들이 일일이 수작업으로 하던 것을 신기술을 활용하여 자동화한 시스템이었기 때문에, 도입 당시 호주 정부는 매우 획기적인 정부혁신으로 홍보하였다.[2] 그러나 〈표 1〉의 주요 연혁에서 확인할 수 있듯이 이 시스템은 도입 이후 수많은 비판을 받았다. 본격적으로 이러한 비판들에 대해서 검토하기 이전에 호주 정부가 지급하는 복지수당과 로보데트 시스템의 도입배경에 대하여 살펴본다.

2) https://www.paulramsayfoundation.org.au/news-resources/beyond-robodebt-spotlight-on-economic-justice-australia (접속일 : 2024.3.5.)

〈표 1〉 로보데트 스캔들 관련 주요 연혁

시기	주요 내용
2014.6월	"신뢰할만한 데이터기반 평가" 개념에 관한 복지부(DHS) 내부의 문서작성
2015년	로보데트 시스템 파일럿 프로그램 시행
2015.3월	내각 예산심의위원회에 로보데트 관련 정책 제시
2015.5월	2015-16년 정부예산안 발표
2016.7월	로보데트 시스템 시행
2017.4월	연방옴부즈만 보고서 발행
2017년	상원청문회 진행
2019.2월	연방법원 제소(Masterton test case)
2019.3월	법제처에서 복지부에 권고안
2019년	제2차 상원청문회 진행
2019.9월	법무법인 고든에서 집단소송 제기할 것을 발표, 법제처장 권고안
2019.11월	연방법원, 소득평균화가 불법적이라고 판결 장관이 총리에게 법제처 권고안 전달
2019.11.19	복지수당 초과지급분에 대한 환수조치 철회
2020.6.20	로보데트 시스템 공식적으로 폐지

출처: https://robodebt.royalcommission.gov.au/system/files/2023-09/rrc-accessible-full-report.PDF

1. 호주 사회서비스청의 복지수당과 로보데트 시스템의 도입

호주의 사회서비스청(Services Australia)[3]은 호주 복지부 산하기관으로서 호주인 및 호주 가정을 대상으로 각종 복지수당, 의료보험, 아동수당, 재해구호수당 등의 지급을 총괄 담당하는 기관이다.[4] 2023년 기준 호주 전역에 걸쳐 318개의 사무소가 개설되어 있으며, 연 2,200억 호주달러(약 200조 원)의 복지수당을 지급하고 있다.

2024년 7월 현재, 사회서비스청이 취급하는 복지수당 지급대상은 크게 가정, 노인, 장애인, 돌봄인, 청소년 및 학생, 취업준비자, 농업인 등 7개 분야로 구분되어 있다. 그 외에도 재해수당과 같은 특별수당, 에너지보조금, 격오지보조금, 월세보조금 등 각종 보조수당을 지급하고 있다.[5] 〈표 2〉에서 확인할 수 있는 것처럼 7개 분야에서 호

3) 로보데트 시스템 도입 당시의 기관명은 센터링크(Centrelink)였다.

4) 영주권자들도 대부분의 복지수당 지급대상에 해당되나, 일부 수당의 경우 영주요건이나 대기시간이 필요할 수 있다.

5) 로보데트 시스템이 도입된 2016년의 복지수당 유형 및 지급액과 2024년 현재의 유형 및

주 정부가 지급하는 각종 복지수당의 유형은 총 28개에 이른다. 28개 복지수당의 수급자 해당여부를 판단하는 데에는 소득이나 자산수준, 연령, 가족관계, 거주지, 장애여부, 돌봄참여여부, 학업형태, 취업여부 등 매우 다양한 기준이 복잡하게 활용되고 있다. 예를 들어, 가정에 지급되는 아동돌봄보조금(child care subsidy)이나 가족수당(family tax benefit), 청소년 및 학생에게 지급되는 청소년수당(youth allowance)과 청년학비보조(Austudy), 장애인에게 지급되는 장애인보조연금(disability support pension) 등은 소득과 자산을 고려하여 차등지급된다. 장애인에게 지급되는 복지수당이더라도 장애인이동수당(mobility allowance)과 청소년장애인보조금(child disability assistance payment)은 소득 및 자산과 관계없이 지급된다. 실업 중인 취업준비생에게 지급되는 실업수당(jobseeker payment)은 결혼 및 자녀유무, 연령, 한부모가정 여부 등에 따라 지급액이 상이하다.

개인이 받는 연간 수령액과 지급빈도도 개인별 편차가 매우 크다. 1회성 보조금인 대학접근보조금(tertiary access payment)의 대상자에게는 5,000 호주달러가 1회 지급되지만, 아동돌봄보조금이나 가족수당 등은 연중 내내 지급될 수도 있다. 육아휴직수당(parental leave pay)이나 육아수당(parenting payment), 실업수당(jobseeker payment)은 개인의 사유가 발생하는 특정 시기 동안에만 지급이 된다. 따라서 복지수당의 연간 수급액은 개인에 따라 적게는 수백 달러에서 많게는 수만 달러에 이른다.

이렇듯 기준이 매우 복합적이고 개인별 편차가 크며 복지수당의 유형에 따라 수급대상자의 수가 최대 수십만 명 이상이 될 수 있다는 점을 감안한다면 개개인별로 기준충족여부를 확인하는 데 엄청난 인력과 자원이 소요될 수밖에 없다. 로보데트 시스템이 도입되기 전에는 개인별 기준 충족여부 판단과 정확한 복지수당 산정에 수많은 인력과 시간이 투입되었다. 2015년 발표된 호주 정부 예산안에 의하면 이러한 수작업에 들어간 예산이 5년 간 약 17억 호주달러(한화 약 1.5조 원)에 달하였다(Commonwealth of Australia, 2023: xxiii). 천문학적 재정이 투입됨에도 불구하고 복지수당 지급이 정확하게 이뤄지고 있는지는 여전히 불확실하였다. 2013년 9월 총선에서 자유국민연합(Liberal-National Coalition)의 토니 애벗(Tony Abbot) 대표는 노동당 정부 하에

지급액은 동일하지 않다. 복지수당의 유형은 한 해에도 수시로 바뀌어 왔기 때문에 그 역사적 변천을 요약하는 것은 어려운 일이다. 유형과 지급액에 있어 시대적 차이가 존재하긴 하지만, 현재의 복지수당 지급규모 등을 제공하는 것이 호주 정부의 복지수당 지급 개요를 확인하는 데 유용하다는 판단하에 본 논문에서는 가장 최근 시점의 제도를 확인하여 소개하고자 한다. 호주 정부 복지수당 유형의 역사적 변천에 대해서는 다음을 참조하시오.
https://www.servicesaustralia.gov.au/sites/default/files/documents/co029-1509en.pdf (접속일 : 2024.6.1.)

서 방만한 지출과 재정낭비가 이뤄졌다고 주장하면서 예산통제와 정부부채 감소를 핵심공약으로 걸고 승리하였다. 이후 2014년 7월 사회서비스부는 각종 복지수당을 과다 지급받아 정부에 부채를 지고 있는 사람들을 대상으로 부채를 환수하는 범정부적 전략을 수립하였고, 2015년 2월 사회서비스부 장관에 임명된 스콧 모리슨 장관[6]은 "호주인들은 사회보장시스템을 악용하려는 사람들을 용납하지 않을 것"이라면서 강력한 복지감시제도의 도입을 천명하였다(Commonwealth of Australia, 2023: xxiii). 이후 시스템의 도입은 일사천리로 진행되었다. 2015년 5월, 차년도 정부예산안을 발표하면서 호주 정부는 사회보장급여 지급과 관련하여 사기방지 및 부채회수를 강화하기 위한 평가절차를 개선하겠다고 발표하였고, 이는 복지수당지급시스템의 완결성(integrity)을 제고한다는 미명 하에 2016년 7월 로보데트 시스템의 전격 도입으로 이어졌다.

〈표 2〉 호주의 복지수당 종류와 수령액(2024.6월 기준)

대상	수당 유형	내용	개인 수령액 (호주달러)
가족	아동돌봄보조금	보육원 이용비용에 대한 보조금	2주에 약 1,000불
	추가 아동돌봄보조금	아동돌봄보조금 수령자에 대한 추가지원	개인상황에 따라 매우 상이
	고아돌봄연금	고아의 돌봄을 책임지는 이에 대한 지원	2주에 최대 77불
	가족수당	육아비용에 대한 지원금	아동 1인당 최대 570불
	출산지원금 및 출산축하금	신생아 출산이나 입양으로 인한 1회성 지원금 및 가족수당에 대한 추가지원금	아동 1인당 667불, 13주간 2,000불
	육아휴직수당	육아휴직 중인 부모를 위한 보조금	1주에 915불
	육아수당	유아를 책임지는 이에 대한 수입보조금	2주에 최대 987불
	한부모가족 보조금	한부모가정에게 지급되는 수당	연 300불
	유산수당	임신 중 아이 유산시 지급되는 일회성 수당	4,225불
노인	노령연금	노령연금 수령연령에 도달한 노인에게 지급하는 연금	2주에 최대 1,116불
	주택연금	본인이 소유한 주택을 담보로 수령할 수 있는 연금	주택가격에 따라 상이
장애인	장애인보조연금	2년 이상 지속되는 장애를 가진 이에게 지급	개인상황에 따라 매우 상이
	필수의료기기수당	필수적인 의료기기 유지를 위해 필요한 전기세 보조	기기당 연 191불
	장애인이동수당	장애인의 근로, 학업, 취업 등을 위해 소요되는 여비지원	2주에 115불
	청소년장애인보조금	소득보조를 받는 장애청소년에 대한 추가지원금	2주에 153불

6) 2018년부터 2022년까지 총리로 재임하였다.

대상	수당 유형	내용	개인 수령액 (호주달러)
돌봄인	돌봄인보조금	일상적 돌봄을 제공하는 이에 대한 추가지원금	2주에 153불
	돌봄인수당	6개월 이상의 돌봄을 필요로 하는 이에게 지속적 돌봄을 제공하는 돌봄제공자에 대한 지원금	2주에 최대 1,116불
	돌봄인지원금	돌봄인보조금이나 돌봄인수당 수령자에 대한 추가적인 연간지원금	연 600불
	아동장애인보조수당	돌봄인보조금 수령자가 장애아동을 돌볼 경우의 추가 연간지원금	아동 1인당 연 1,000불
청소년 및 학생	원주민장학제도	호주 원주민에 대한 학비 (생활비, 교통비 등 포함) 지원	석박사 과정생은 2주에 1,231불
	격리된 아동보조제도	격오지 거주, 장애 등의 이유로 학교에 다니지 못하는 학생의 부모/보호자에게 지급되는 지원금	연 최대 12,816불
	청년학비보조	25세 이상의 대학생/견습생에게 지급되는 학비보조금	2주에 최대 690불
	연금소득자교육지원금	특정 복지수당 수급자들의 학업지원을 위한 보조금	2주에 최대 62불
	대학접근보조금	집에서 멀리 떨어진 대학에 입학한 학생들에게 지급되는 1회성 보조금	1회 5,000불
	청소년수당(학생)	24세 이하의 학생/견습생에게 지급되는 수당	2주에 최대 690불
취업준비자	실업수당	22세 이상 연금수급연령 이하의 국민에게 지급되는 실업수당	2주에 최대 987불
	청소년수당(취업준비생)	21세 이하의 취업준비생/실업자에게 지급되는 수당	2주에 최대 987불
농업인	농가지원수당	재정적 어려움을 겪는 농가에 지원되는 보조금	2주에 최대 987불

출처: https://www.servicesaustralia.gov.au/guide-to-australian-government-payments을 재구성 (접속일: 2024.6.18.)

* 상기 유형과 수령액은 2024년 6월 기준으로서, 로보데트 시스템 도입 당시와는 유형과 금액에서 차이가 존재함. 또한 대부분의 경우 개인상황과 소득/재산 수준에 따라 실제 수령금액에서 차이가 발생함.

2. 데이터매칭과 소득평균화 : 잘못된 전제에서 나온 잘못된 부채환수통보

복지수급대상자를 특정하고 이들에게 지급되는 복지수당을 정확하게 산정하기 위하여 호주 정부는 두 가지 경로로 파악한 소득데이터를 활용하였다. 하나는 복지부 산하 센터링크(Centrelink)에 복지수급대상자들이 신고하는 월소득 데이터, 보다 정확히는 2주간 소득 데이터이고,[7] 다른 하나는 국세청에서 수집하는 납세자들의 연간 개인소득

7) 복지수당 지급은 기본적으로 소득기준으로 이루어지기 때문에, 호주의 복지수당 지급을 담당하는 센터링크는 복지수급대상자들로부터 2주간의 소득에 대해서 리포트할 것을 요구해왔다.

(PayG) 데이터다. 로보테트 시스템은 데이터매칭 알고리즘을 통해 두 데이터를 매칭하여 복지수당 과다수급자를 가려내고 이들에게 부채환수통지를 하는 시스템이었다(Braithwaite, 2020).

로보데트 시스템은 2016년에 본격 도입되었지만, 호주 정부는 2010년의 복지수당 지급액부터 검토하기 시작하였다. 데이터매칭을 통하여 2010년부터 2012년까지 총 3년간 정부가 지급한 복지수당 지급건수 중 본인이 센터링크에 신고한 연간수입과 국세청이 파악한 소득이 불일치하는 866,857건을 가려내었고, 해당 복지수급자들에게 이러한 불일치를 설명할 것을 요구하였다(Commonwealth of Australia, 2023: xxiv). 수급자 개인이 불일치의 이유를 설명하지 못하거나 국세청 소득데이터에 동의하는 경우, 국세청 소득데이터를 평균화하여 개인의 2주간 소득 평균액이 산정되었다. 즉, 복지수급대상자들이 센터링크에 신고한 2주간 소득 데이터의 정확성을 확인하기 위하여 국세청으로부터 입수한 연소득 데이터가 주간 평균화되었고, 연소득 데이터를 26으로 나누어 2주간 소득 데이터의 평균값을 구한 후 이를 센터링크에 보고된 2주간 소득 데이터와 비교하여 복지수급자 개인이 정부에 진 부채액을 구한 것이다.

로보데트 시스템에 대한 가장 주요한 비판은 이렇게 계산된 부채환수통보(debt notice)가 잘못된 계산일 수 있다는 점이다. 실업수당의 경우를 예로 들어보자. 실업수당 대상자 A씨는 1년간 4주를 일하였고, 일하는 동안 시급 25달러를 받아 총 4,000달러의 임금소득이 발생(25달러×주40시간×4주)하였다고 가정해보자. A씨는 소득이 발생하지 않은 48주간은 2주에 1,500달러의 실업수당을 지급받았으며, 임금소득이 발생한 4주 간은 삭감된 금액으로 실업수당을 지급받았다. 로보데트 시스템이 도입되기 이전에는 소득이 발생한 4주간 50%가 삭감된 1,500달러만을 수령하게 되어 1년 동안 받아야 할 실업수당 중 1,500달러를 덜 받았다. 그러나, 로보데트 시스템에 의하면 4,000달러의 임금소득이 1년 평균소득으로 인정되므로, A씨는 매주 38.5달러의 실업수당을 삭감받아 연간 약 2,000달러의 실업수당을 덜 받아야 한다(38.5달러×52주). 이 경우, 로보데트 시스템은 A씨에게 500달러의 부채를 정부에 상환해야 함을 자동으로 통보하게 된다.

소득 평균화(income averaging)가 적용되기 때문에, 로보데트 시스템의 데이터매칭 알고리즘은 센터링크에서 수집한 2주간 소득과 국세청의 연소득 데이터를 평균화한 2주간 소득 간에, 그리고 후자에 근거한 복지수당 산정에 있어 편차를 인지할 수밖에 없다. 이는 개인소득이 연중 고르게 발생한다면 타당한 전제라고 할 수도 있으나, 비정규직이나 일용직에 종사하고 있어서 수입이 불규칙적으로 발생하는 다수의 복지수급대상자들에게는 적용되기 어렵다(O'Donovan 2019). 잘못 통보된 부채환수금액 총액

이 얼마인지에 관해서는 정확한 통계가 나와 있지 않으나, 가장 최근의 보고에 의하면 약 47만 건이 잘못 통보되었고 총액은 약 10억 호주달러에 이른다(Henriques-Gomes 2020). 그러나, 전체 대비 잘못 통보된 건의 비중이 얼마나 되는지, 이들에게 왜 통보가 잘못되었는지 등에 대해서는 아직까지 구체적으로 밝혀지지 않았다.

3. 기술에 대한 과의존과 정부의 책임전가

호주 복지부와 센터링크가 로보데트 시스템이 가지고 있는 이러한 편차를 사후적으로 교정하려는 노력을 하지 않으면서 문제는 심각해졌다. 차액이 발생하여 복지수급자가 수입에 비해 과도한 복지수당과 급여를 받았다고 시스템이 판단을 하게 되면, 이들에게 부채환수통지서가 자동적으로 발급되었다. 정부가 시스템의 오류가능성에 대해서 사후적으로 검토하지 않으면서, 이를 증명해야 할 책임은 센터링크로부터 부채환수통보를 받은 대상자들에게 전가될 수 밖에 없었다. 이들에게 주어진 선택지는 통보받은 금액을 정부에 갚거나, 부채환수통보가 틀렸다는 점을 스스로 증명하는 길 뿐이었다.

많은 복지수급자들은 부채에 관한 잘못된 통보를 받고 정부에 이를 반드시 갚아야 한다는 압박감에 시달리게 되었다. 부채를 갚지 않으면 개인의 신용도가 저하되는 결과를 초래할 수 있었다. 개인이 갚아야 할 부채액 산정의 근거에 대한 정보가 부재한 상황에서 복지수급자들이 정부를 대상으로 소송을 진행하는 것은 현실적으로 매우 어려운 일이었다. 2022년 8월부터 2023년 6월까지 약 10개월 간 진행된 왕립조사위원회(Royal Commission into the Robodebt Scheme)는 다수의 이의제기가 행정심판소(Administrative Appeals Tribunal)에 제소되었고 정부가 잘못하였다는 판결이 여러 번 나왔다는 점을 사후적으로 밝혀내었다. 그러나, 호주의 행정심판소는 개인적인 송사를 다루기 때문에 그 결과가 대중에게 공개되지 않는다. 결과가 공개되려면 정부가 항소를 해야 하나, 해당 판결에 대해서 호주 정부는 상급법원에 항소를 진행하지 않았기 때문에 행정심판소의 개별 판결이 대중에게 알려질 수 없었다.[8] 결국 수백 달러에서 수천 달러에 이르는 부채환수통고를 받은 수많은 사회적 약자들은 심리적 불안감과 스트레스, 감정적 트라우마, 수치심 등을 견뎌야 했고, 심지어는 이로 인해 자살하는 경우도 다수 발생하였다.

결국 2020년 호주 정부는 약 47만 건의 부채통보가 잘못 고지되었다는 점을 인정하고 이 시스템을 폐지하겠다고 발표하였다. 이는 스콧 모리슨 자유국민연합 정부가

8) 이후 호주 정부는 해당 판결을 내린 행정심판소 재판관인 테리 카니(Terry Carney) 교수의 재임명을 거부하였다.

총선에서 패배하는 빌미를 제공하였다. 2023년 7월에 발표된 왕립조사위원회의 조사는 자동화된 의사결정에 대한 인간 개입의 중요성을 강조하는 한편, 정부 내의 자동화된 의사결정을 통제할 수 있는 일관적인 법적기틀을 마련할 것을 권고하고 있다.[9)]

Ⅲ. 자동화된 복지행정의 위험성

지금까지 살펴본 것처럼 로보데트 시스템은 그 설계와 운용에 있어서 많은 문제점을 지니고 있었다. 이 사례는 본격적으로 인공지능 시대를 맞아 인공지능에 기반하여 다양한 정부서비스를 디지털 전환하고자 하는 수많은 국가들에게 주요한 시사점을 던져준다.

가장 근본적인 시사점은 인공지능과 같은 신기술의 활용을 통해 행정시스템을 개혁하고 정부의 서비스를 고도화하고자 할 때, 정부서비스의 목적을 호도해서는 안된다는 점이다. 로보데트 스캔들은 자동화된 복지서비스가 인간의 복지향상이 아니라 정부지출 효율화를 목적으로 도입될 경우 발생할 수 있는 해악을 적나라하게 보여주는 사례다. 앞서 호주 자유국민연합 정부의 정책방향에서 언급되었듯이, 로보데트 시스템의 목적은 인간을 위한 복지서비스 향상이 아니라 복지수당 과다지급 방지를 통한 정부지출 효율화였다. 정부시스템의 효율화와 재정건전성 확보는 공공부문에서 인공지능을 도입하는 다수의 정부가 내걸고 있는 핵심적 배경이기도 하다. 호주의 로보데트 스캔들은 신기술의 적용을 통해 행정서비스의 효율화를 도모할 때 인간중심성을 잊지 않아야 한다는 중요한 교훈을 상기시킨다. 인간중심성을 상실한 자동화된 의사결정은 사회의 구성원들에게, 특히 사회적으로 취약한 이들에게 잔인한 결과를 야기할 수 있다. 호주 왕립조사위원회의 보고서는 이러한 반성을 잘 담고 있다.

> 사회복지정책과 관련하여 취할 수 있는 사고방식에는 여러 가지가 있다. 하나는 많은 시민이 인생의 다양한 시기에 소득지원이 필요하다는 점을 인정하는 것이다. 학업이나 구직 과정에서 일시적으로 필요한 경우도 있으며, 연령이나 불우함 또는 장애 등의 이유로 말미암아 소득지원이 보다 장기적으로 필요한 이들도 있다. 이러한 입장에서는 이들에 대한 지원을 기꺼이, 적절하게, 그리고 존중하는 마음으로 제공해야 한다고 믿는다. 다른 접근

9) https://robodebt.royalcommission.gov.au/publications/report (검색일 : 2023.10.26.)

> 방식은 사회보장 혜택을 받는 사람들을 국가경제의 발목을 잡는 존재로 간주하고, 가능한 모든 방법을 동원해 이에 대한 예산을 줄여야 한다고 주장한다. 이러한 입장에서는 복지수급자를 납세자의 부담이라 지칭하고, 복지혜택을 청구한 사람들에게 부담스러운 조건을 요구하고, 공무원의 복지지원을 최소화하고, 정당하든 아니든 혜택을 회수하고, 사회보장수급자의 상태를 불쾌하고 바람직하지 않게 만들고자 한다. 로보데트 시스템은 후자의 대표적 사례다(Commonwealth of Australia, 2023: xxiii).

둘째, 자동화된 행정서비스의 도입은 정부의 진실성(government integrity) 문제를 발생시킬 수 있다. 이 때 진실성은 크게 두 가지의 질문으로 구성될 수 있다. 첫째는 소명성(accountability)의 문제다. 즉, 자동화된 의사결정이 정확한지의 여부를 누가 소명해야 하는가의 문제라고 할 수 있다. 자동화 시스템이 내리는 의사결정에 대한 소명의 책임은 알고리즘도, 그 결정이 적용되는 당사자도 아니라 행정적 결정의 주체인 해당 행정기관에 있어야 한다. 로보데트 시스템에서는 부채환수라는 행정적 결정이 자동화된 시스템에 의하여 이루어졌는데, 그 정확성을 확인하고 소명해야할 책임이 복지수급자 개개인에게 전가된 것이 문제라고 할 수 있다. 두 번째 질문은 제도운영의 책임성(responsibility)과 관련이 있다. 이는 자동화된 행정서비스를 채택함으로써 발생하는 문제에 대해서 누가 책임을 져야 하고, 문제를 인지했을 때의 책임있는 행동은 무엇이냐의 문제로 귀결된다. 호주 정부는 로보데트 시스템으로 인해 문제가 불거졌을 때 이러한 문제점이 발생한 원인을 규명하고 대안을 마련하려 노력하기 보다는 정부차원의 문제로 부각되는 것을 경계하고 시스템에 대한 검토를 외면하였다. 이는 정권의 기반을 흔들었고, 결국 총선에서의 패배와 책임자에 대한 각종 민형사 소송으로 이어지게 되었다.

인공지능 기반 행정서비스의 종류는 다양할 수 있다. 인간의 판단을 보조할 수 있는 정보를 보다 효과적으로 전달하는 서비스일 수도 있고, 행정적 판단을 내리는 서비스일 수도 있다. 행정적 판단을 내릴 때에도 법령에 기속된 행정적 판단을 내리는 서비스와 재량적 판단을 내리는 서비스로 나뉠 수 있다. 단순한 정보제공서비스의 경우에는 행정서비스 사용자나 공급자의 판단을 보조하는 역할에 그치기 때문에, 해당 서비스가 제공하는 정보는 다양하게 접할 수 있는 정보 중 하나로 간주될 것이다. 이 경우 행정서비스와 관련한 정부의 진실성은 큰 문제로 대두되지 않을 수 있다. 그러나, 인공지능에 기반하여 행정적·정책적 판단을 내리는 경우에는 소명성과 책임성의 문제를 반드시 고민할 필요가 있다.

셋째, 로보데트 스캔들은 견제되지 않은(unchecked) 정책적·행정적 결정이 사회적 약자인 개인에 대한 정부의 지배(domination)로 쉽게 발전될 수 있다는 점을 말해준다. 부채환수통보를 받은 복지수급자들에게 선택지가 주어지지 않은 것은 아니다. 이들에게는 부채환수를 거부하거나, 부채환수통보가 부당함을 정부에 소명하거나, 행정심판소 등을 통해 제소할 수 있는 선택지가 주어졌다. 정부를 상대로 행정소송을 제기하는 것 또한 선택지 중 하나가 될 수 있다. 문제는 이러한 선택지들이 실제 개인이 선택할 수 있는 실질적인 선택지가 되기 어려웠다는 점이다. 부채환수통보의 근거에 대한 정보가 부족한 상황에서 정부에 대한 소명은 결정을 번복할만큼 효력이 없었고, 소명되지 않은 부채환수는 개인신용의 저하를 야기할 뿐이었다. 행정심판소를 통한 부당함의 호소는 현실적으로 작동하지 않았다. 부당한 정부 결정에 대한 개인의 저항력(contestatory power)이 발휘될 수 없는 조건 속에서, 잘못된 부채환수통보를 받은 다수의 시민들은 정부에 대한 불신을 갖게 된 것을 넘어 스트레스와 감정적 트라우마, 수치심을 겪을 수 밖에 없었다. 수많은 호주 시민들의 개인적 자유가 침해되고, 정부의 자의적 권력에 종속되는 상황을 맞은 것이다. 로보데트 스캔들은 행정서비스가 지닌 독점적 성격이 견제되지 않을 경우 자의적 지배로 전이될 수 있다는 점을 보여주는 대표적 사례다.

Ⅳ. 위험성 대응을 위한 인공지능 서비스의 조건

1. 위험성 대응을 위한 공공부문의 과제

다수의 기존 연구들은 공공부문에서 인공지능 기술의 활용에 있어 공공성을 중시한다. 즉 공공의 이익을 증진하고 공공의 가치를 제공해야 하는 방향으로 기술의 활용이 이루어질 필요가 있다는 것이다(Cath et al., 2018; Crawford, 2016). 공공부문에서 인공지능 기술 적용의 결정은 민간 부문과는 달리 복잡한 정책, 사회적, 법적, 경제적 요소들을 포함하고 있기 때문이다. 이는 공공부문 인공지능 프로젝트가 가지고 있는 고유한 도전이기도 하다.

이러한 점을 고려한다면, 공공부문에서는 인공지능의 위험을 다룰 때 더욱 엄격한 기준이 필요하다. 공공부문 인공지능 프로젝트는 단순한 효율성 향상을 넘어, 형평성, 공정성, 공익 증진, 취약계층 보호, 인간 존엄성, 프로그램의 경제적 지속 가능성, 법치주의 유지 등 다양한 공공 가치를 고려해야 한다(Osborne, Radnor & Nasi,

2013; Benington & Moore, 2011; Jørgensen & Bozeman, 2007). 특히 투명성과 공정성은 공공부문 인공지능 시스템의 운영과 의사결정에서 중요한 요소이다. 납세자의 자금으로 운영되는 시스템은 정기적인 감시와 감독을 받아야 하기 때문이다(Edwards and Veale, 2017; Bryson and Winfield, 2017; Desouza et al., 2020). 또한, 공공부문에서 잘못된 결정을 내릴 위험이 높아질수록 데이터와 알고리즘의 평가는 더욱 중요해진다. 민간 부문에서는 가치 대비 위험의 균형을 고려하여 의사결정을 내리곤 하지만, 공공부문에서의 위험평가는 근본적으로 다를 수밖에 없다. 요약하자면, 공공부문 인공지능 사용은 공익을 증진하고, 시민의 권익을 보호하며, 사회 전체에 긍정적인 영향을 미칠 수 있도록 보장되어야 한다.

반면, 공공부문에서 인공지능 기술을 활용하는 데 있어 엄격한 기준은 민간 부문에 비해 혁신을 경직되게 만들 수 있다. 공공부문 인공지능 프로젝트에 요구되는 높은 기준과 감시수준이 지니는 경직성으로 인해 공공부문이 민간 부문의 혁신 속도를 따라가지 못하는 결과를 초래할 가능성은 상존한다. 따라서, 이러한 엄격한 기준이 혁신을 저해하지 않으면서도 공공 서비스의 품질을 유지할 수 있는 방법을 모색하는 것이 중요하다.

선행연구에서는 공공부문에 요구되는 엄격한 기준을 만족시키면서도 인공지능의 도입을 촉진하기 위해 다양한 접근법을 제안하고 있다. 크게 네 가지로 정리할 수 있다. 첫째, 기술적 도구를 활용한 접근이다. 예를 들어, 2018년 뉴욕에 본사를 둔 한 인공지능 스타트업은 알고리즘이 특정 통계나 특성을 편향적으로 처리하는지 여부를 확인하기 위해 감사 도구를 오픈 소스로 공개했다(Johnson, 2018). 이러한 도구는 공공부문에서도 편향성을 최소화하고 투명성을 높이는 좋은 사례가 될 수 있다.

둘째, 정기적인 모니터링과 감사 등을 통해 인공지능 시스템의 윤리적 기준을 유지할 수 있다. 이러한 절차를 통해 기관은 다양한 계층의 개인을 차별하지 않는다는 것을 공개적으로 증명할 수 있으며, 높은 윤리적 기준을 준수하고 있다는 점을 대내외적으로 알릴 수 있다. 이는 공공부문에서도 투명성과 신뢰성을 높이는 데 기여할 수 있다.

셋째, 외주화를 통한 접근도 고려할 만하다. 공공기관은 위험과 가치를 평가하여 파트너십을 형성할 수 있다. 내부에서 개발할 전문 지식과 외부에서 소싱할 수 있는 지식을 신중하게 구분하여 결정하는 것이 필요하다. 이는 공공부문이 빠르게 변화하는 기술 환경에 적응하면서도 비용 효율성을 유지할 수 있는 방법이다.

넷째, 공공부문 내에 인공지능 기술에 대한 충분한 이해와 역량을 갖추는 것이 중요하다. 공공기관은 다음 세대의 시스템을 설계하고 배포된 시스템을 조사, 수정, 학습할 수 있는 기술 역량을 유지해야 한다. 이는 인공지능 기술의 지속적인 발전과 함께

공공부문이 혁신을 지속적으로 수용할 수 있도록 보장한다. 공공부문이 인공지능 역량을 갖추게 되면 기술 도입 과정에서 발생할 수 있는 오류를 줄이고, 투명하고 공정한 인공지능 시스템을 운영하는 데 도움이 된다(Goodman & Flaxman, 2017).

2. 위험성 대응을 위한 인간의 개입, HITL AI

이상의 논의를 정리하면, 공공부문에서 인공지능의 위험성에 대응하면서 인공지능 서비스를 도입하기 위해서는 인공지능 시스템이 개발되고 서비스가 제공되는 과정이 공정하게 이루어지고 있는지를 투명하게 감시하고, 잘못된(또는 잘못될 수 있는) 알고리즘을 올바르게 수정할 수 있는 개념적이면서도 실질적인 체계가 요구된다 하겠다. 이때 감시와 수정은 인공지능이 아닌 인간의 개입(외주화된 인력이건 내부의 인력이건 간에)에 의해 이루어져야 한다. 따라서, 인공지능의 위험성에 대응하기 위해 알고리즘에 인간을 포함시키는 방법이 점점 중요해지고 있다. 선행연구에서는 이러한 접근법을 Human-in-the-loop (HITL) ML, 또는 HITL AI라고 정의하고 있다(Edwards & Veale, 2017; Bryson & Winfield, 2017). 이 접근 방식은 인간의 전문 지식을 활용하여 모델의 정확성과 효율성을 높이는 것을 목표로 한다(Amershi et al., 2014; Wang, Guo, & Chen, 2022).

HITL ML은 Human-Assisted Machine Learning(HAML)과 Machine-Learning-Assisted Human(MLAH)으로도 분류될 수 있다(Monarch, 2021). HAML은 인간의 피드백을 통해 머신러닝 알고리즘의 성능을 개선하는 것을 의미한다. 이는 라벨링된 데이터를 제공하거나, 모델의 출력물에 대한 피드백을 제공하여 모델이 학습하고 개선할 수 있도록 돕는다. 반면, MLAH는 머신러닝 알고리즘을 사용하여 인간이 작업을 수행하는 데 도움을 주는 방식을 의미한다. 예를 들어, 데이터 전처리 작업을 자동화하거나 인간 주석자에게 제안을 함으로써 작업 효율성과 정확성을 높이는 것이다. 본 연구에서는 인간의 개입으로 머신러닝 알고리즘을 개선하는 HAML을 통한 인공지능 위험성 대응에 초점을 맞추고자 한다.

HITL ML에는 학습 과정의 주도자에 따라 여러 가지 접근법이 있다(Holmberg et al., 2020). 대표적인 접근법은 다음과 같다. 첫째는 Active Learning(AL)이다. 이는 기계 학습 모델이 가장 유용한 데이터를 선택하고, 인간에게 해당 데이터를 라벨링하도록 요청하는 방식이다(Settles, 2009). AL은 데이터 라벨링의 효율성을 높이고, 모델의 성능을 빠르게 개선할 수 있다. 둘째는 Interactive Machine Learning(IML)이다. IML은 사용자가 학습 알고리즘과 상호작용하며 유용한 결과물을 생성하는 과정이다

(Amershi et al., 2014). 사용자는 반복적인 피드백을 통해 모델을 지속적으로 개선할 수 있다. 이 과정에서 ML 전문가, 데이터 과학자, 크라우드소싱 작업자, 도메인 전문가 등 다양한 역할을 수행할 수 있다. 셋째, Machine Teaching(MT)이다. MT는 도메인 전문가가 학습 알고리즘에 직접 가르치는 방식으로, 모델의 성능을 향상시키는 데 중요한 역할을 한다(Simard et al., 2017). MT는 ML 알고리즘을 교체 가능한 구성 요소로 취급하며, 이는 큰 패러다임 전환을 의미한다.

3. 인간의 3중 개입(triple-loop human intervention) 개념 제안

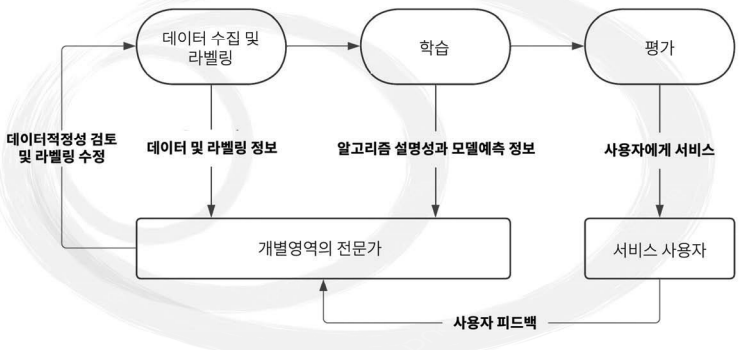
HITL AI 기법은 데이터 수집 및 라벨링, 모델 학습, 모델 평가 등 다양한 단계에 적용될 수 있다. 본 논문에서는 공공부문에서 인공지능 서비스 도입을 위해 인공지능 위험성에 대응하기 위한 인간의 개입이 중첩되어서 작동해야 할 필요가 있다고 주장하고자 한다. 보다 구체적으로는 인간의 3중 개입(triple-loop human intervention) 개념이다.

첫 번째 인간의 개입은 데이터 수집 및 라벨링 과정에서의 개입이다. 이 과정에서 특정 도메인 영역의 인간 전문가는 인공지능이 학습할 데이터의 품질을 보장하고 라벨이 정확하게 지정되었는지를 검토한다. 인간 검토자는 수집된 데이터가 편향되지는 않았는지, 인공지능 서비스의 목적에 부합한 데이터인지, 신뢰할 만한 출처를 가진 데이터인지 등을 검토함으로써 데이터의 품질을 보장할 수 있다. 데이터 라벨링의 검토는 특히 복잡하거나 미묘한 차이를 포함하는 데이터셋에서 매우 중요하다. 예를 들어, 의료 영상 데이터의 경우 의사가 이미지의 특정 부분을 라벨링하여 학습 데이터를 준비한다. 이는 알고리즘이 보다 정확한 예측을 할 수 있도록 도와준다.

두 번째 개입은 모델 학습 과정에서의 개입이다. 이 과정에서는 초기 모델이 학습되고 나서 인간 검토자가 모델의 출력물을 검토한다. 이때 모델의 출력물이란 모델의 알고리즘을 설명할 수 있는 데이터와 모델의 예측(또는 생성) 데이터이다. 모델의 알고리즘을 설명할 수 있는 데이터는 다양하다. 예를 들어, 거대언어모델(LLM)을 파인튜닝하여 감정분류모델을 만들었다고 하면, 이때 '긍정'과 '부정'으로 분류하는데 어떤 단어들이 높은 가중치를 부여받았는지를 출력하게 할 수 있다. 인간 검토자는 출력된 단어들을 통해 분류모델이 올바르게 학습했는지를 검토할 수 있다. 이후, 특정 단어가 포함된 데이터를 출력하여 라벨링이 제대로 되었는지를 확인할 수도 있고, 해당 단어의 가중치를 직접 조정할 수도 있다. 이러한 피드백은 모델이 잘못된 예측을 수정하고, 새로운 패턴을 학습하며, 전반적인 성능을 개선하는 데 사용된다(Amershi et al.,

2014).

세 번째 단계는 모델 평가 단계에서의 개입이다. 이때, 인간 검토자가 모델의 평가에 직접 개입하는 것보다는 인공지능 서비스 사용자의 피드백을 수집하여 검토하는 것이 보다 효과적이다. 검토자가 검토의 목적으로 서비스를 평가하는 것은 한계가 있으며, 오히려 다수의 서비스 이용자가 다양한 형태로 서비스를 이용한 결과를 통해 예기치 않았던 문제들을 발견할 수 있기 때문이다. 여기서 인간 검토자는 사용자 피드백을 통해 모델이 실제 환경에서 의도한 대로(설계한 대로) 작동하는지 평가하고, 필요한 경우 추가적인 조정을 수행하도록 요구할 수 있다(Fails & Olsen, 2003). 많은 민간 인공지능 서비스에서는 인공지능 시스템의 고도화를 위해 사용자 피드백 기반 강화학습(Reinforced Learning with Human Feedback, RLHF)을 수행한다. 이때, 사용자는 만족(thumb up) 또는 불만족(thumb down) 버튼을 통해 서비스의 만족도를 표현하게 된다. 강화학습의 알고리즘은 사람들이 만족한 것과 유사한 결과의 생성을 강화하는 것이기에, 민간의 인공지능 서비스 고도화에는 매우 적합한 방식이다. 그러나 전술한 바와 같이 공공부분의 인공지능 서비스는 사용자의 만족 여부뿐만 아니라 공공의 이익과 가치를 중요시하기 때문에 인간 검토자가 사용자의 만족/불만족의 이유와 맥락을 파악할 수 있어야 한다. 따라서 보다 구체적인 피드백 정보를 수집하고 이를 모델에 반영할 수 있는 방안에 대한 고려가 필요하다.



〈그림 1〉 공공부문 인공지능 위험성 대응을 위한 인간의 3중 개입 (triple-loop human intervention) 개념

V. 결론

행정서비스의 디지털 전환은 효율성을 높이고 다양한 혜택을 제공할 수 있지만, 그 과정에서 발생할 수 있는 다양한 위험성을 간과해서는 안 된다. 특히 복지행정과 같은 민감한 분야에서 자동화된 시스템에 의존할 경우 사회적 약자에게 심각한 영향을 미칠 수 있다. 호주의 로보데트 스캔들은 이러한 위험을 여실히 보여주는 사례다. 이 사례에서는 복지수당 초과지급분을 자동화 시스템을 통해 계산하고 환수 조치를 취하면서 많은 사회적 약자들이 잘못된 통지를 받았고, 이로 인해 심리적, 경제적 어려움을 겪었다. 이는 인공지능 기반 행정서비스 도입 시 인간중심적 접근이 필수적임을 시사한다.

인공지능 서비스의 도입과 운영에서 발생할 수 있는 위험성을 최소화하고 인간중심적인 접근을 가능하게 하기 위해서는 다음과 같은 조건들을 충족시켜야 한다. 첫째, 투명성과 공정성이 확보되어야 한다. 인공지능 시스템의 설계 및 운영 과정은 투명하고 공정해야 한다. 이는 시스템이 어떠한 방식으로 작동하는지, 데이터가 어떻게 처리되고 있는지, 그리고 의사결정이 어떻게 이루어지는지를 명확하게 공개하는 것을 의미한다. 시스템의 작동을 모니터링하고, 오류가 발생할 경우 이를 신속히 수정할 수 있는 체계가 필수적이다. 이를 위해 정기적인 감사와 평가가 필요하며, 외부 전문가의 검토를 통해 시스템의 투명성을 높여야 한다. 또한, 알고리즘이 편향되지 않도록 다양한 데이터를 사용하고, 의사결정 과정에서 특정 집단에게 불이익이 가지 않도록 공정성을 유지해야 한다.

둘째, 인공지능의 알고리즘과 의사결정 과정에 인간의 개입이 가능해야 한다. 이는 중요한 의사결정 과정뿐만 아니라, 알고리즘이 형성되는 과정에도 인간의 개입이 필요하다는 것을 의미한다. 예를 들어, 복지수당의 지급 여부를 결정하는 과정에서 자동화된 시스템이 제시한 결과를 최종적으로 확인하고 승인하는 절차를 두는 것을 넘어, 결정이 이루어지는 논리에 대한 검토의 절차를 두어야 한다. 인공지능에 기반한 판단이 오류가 있을 수 있음을 인지하고 영향을 받는 이들의 불만제기에 귀를 기울일 수 있는 제도적 장치를 만들어둘 필요가 있다. 이는 시스템의 오류를 최소화하고, 사회적 약자를 보호하기 위한 중요한 장치이다. 또한, 인간의 개입을 통해 시스템의 윤리적 문제를 감시하고, 인공지능의 판단이 도덕적, 사회적 기준을 충족하는지 확인하는 것이 반드시 필요하다.

셋째, 데이터 품질 관리가 중요하다. 인공지능 시스템의 성능은 입력되는 데이터의 품질에 크게 의존한다. 따라서, 데이터의 정확성, 완전성, 일관성을 보장하기 위한 철저한 관리가 필요하다. 잘못된 데이터는 잘못된 결과를 초래할 수 있으며, 이는 복지수급자들에게 심각한 피해를 줄 수 있다. 데이터 수집 단계에서부터 정제, 저장, 활용

에 이르기까지 데이터 품질 관리 절차를 엄격히 준수해야 한다. 이를 위해 데이터 수집 과정에서 발생할 수 있는 오류를 최소화하고, 수집된 데이터의 출처와 신뢰성을 검증하는 작업이 중요하다. 또한, 데이터의 주기적인 업데이트와 검증을 통해 최신성과 정확성을 유지해야 한다.

마지막으로, 사회적 가치가 고려되어야 한다. 인공지능 시스템의 도입과 운영은 기술적 효율성만을 추구해서는 안 되며, 사회적 가치와 공공의 이익을 우선적으로 고려해야 한다. 이는 특히 취약계층 보호와 사회적 약자를 고려한 정책적 결정이 중요함을 의미한다. 시스템이 제공하는 서비스가 모든 사람에게 공평하게 혜택을 제공할 수 있도록 설계되어야 하며, 사회적 약자를 지원하고 보호하는 기능이 포함되어야 한다. 이를 위해 인공지능 시스템의 개발 단계에서부터 다양한 이해관계자들의 의견을 반영하고, 공공의 이익을 극대화할 수 있는 방안을 모색해야 한다.

본 연구에서는 공공 부문에서 인공지능 도입으로 인한 자동화의 폐해를 최소화하고, 인간중심적인 인공지능 도입을 위해 데이터 수집 및 라벨링, 모델 학습, 모델 평가의 세 단계에서 이루어지는 인간의 3중 개입의 중요성을 강조하였다. 데이터 수집 및 라벨링 단계에서의 인간 개입은 데이터의 품질을 보장함으로써 편향되거나 거짓된 정보를 인공지능이 학습하는 것을 방지하여 데이터의 정확성과 신뢰성을 확보한다. 모델 학습 단계에서의 인간 개입은 모델의 예측 결과를 인간이 검토하고 필요시 조정함으로써 모델의 성능을 지속적으로 개선한다. 모델 평가 단계에서의 인간 개입은 실제 사용자의 피드백을 수집하고 이를 기반으로 모델을 지속적으로 개선하는 역할을 한다. 이러한 3중 개입은 인공지능 시스템이 보다 공정하고 정확하게 작동하도록 보장하는 최소한의 필수적인 절차가 되어야 한다.

디지털 전환의 과정에서 기술적 효율성과 인간 존엄성의 균형을 맞추는 것이 중요하다. 기술은 인간의 삶을 개선하기 위한 도구일 뿐, 그 자체가 목적이 되어서는 안된다. 인공지능의 도입으로 행정서비스의 다양한 부분을 자동화하고 효율화할 수 있지만, 그 목적은 해당 행정서비스의 목적을 보다 잘 달성하기 위한 것이 되어야 한다. 정부지출의 효율화나 복지수당 부당수급자의 판별과 같은 목적이 인공지능 도입의 목적이 되어서는 안된다는 것이다. 정책적 감독과 규제는 자동화된 행정 시스템이 공정하고 투명하게 운영되도록 보장하는 필수적인 요소이다. 이를 통해 시스템의 오류를 발견하고 수정할 수 있는 지속적인 감시와 평가 체계를 마련해야 한다. 특히, 사회적 약자를 보호하는 정책적 노력이 중요하다. 이는 기술적 효율성보다 더 중요한 목표로 설정되어야 한다. 호주의 로보데트 스캔들이 주는 교훈은 디지털 전환이 단순한 기술 혁신을 넘어 인간중심적 가치와 사회적 책임을 고려해야 한다는 점이다.

▣ 참고문헌

김길수. 2019. “공공부문에서 인공지능 활용에 관한 연구.” ≪한국자치행정학보≫, 33(1):27-48.

김대인. 2020. “인공지능과 행정법총론체계의 변화.” ≪경제규제와 법≫, 13(2): 108-128.

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

김이인정・이종훈・황하・홍승헌. 2022. “레그테크 발전 동향과 규제혁신에의 함의: 규제준수 및 리스크 관리를 중심으로.” ≪규제연구≫, 31(2):161-192.

김한솔・김병조. 2023. “국내 공공부문 AI 활용에 관한 탐색적 연구: 기술의 진화(進化)와 사용의 편중(偏重).” ≪한국행정학보≫, 57(2):205-246.

김휘식. 2021. “인공지능에 의한 행정상 자동결정에 대한 규율과 권리구제.” ≪IT와 법 연구≫, 22:301-340.

박동아. 2017. “인공지능 기반 대화형 공공 행정 챗봇 서비스에 관한 연구.” ≪멀티미디어학회논문지≫, 20(8):1347-1356.

박석희・조강주. 2021. “공공기관 사회적 책임 이행수준의 격차 원인 탐색.” ≪한국사회와 행정연구≫, 32(3):29-57.

송진순. 2021. “공공 소통 증진을 위한 인공지능 챗봇의 기능 강화 방안과 정책 제언.” ≪입법과 정책≫, 13(1):339-366.

안준모. 2021. “인공지능을 통한 행정의 고도화: 기회와 도전.” ≪한국행정연구≫, 30(2):1-33.

유순덕. 2023. “인공지능 서비스 영향성 평가를 위한 분석 기준 연구.” ≪The Journal of the Institute of Internet, Broadcasting and Communication: JIIBC≫, 23(1):7-13.

윤상오. 2018. “인공지능 기반 공공서비스의 주요 쟁점에 관한 연구: 챗봇(ChatBot) 서비스를 중심으로.” ≪한국공공관리학보≫, 32(2):83-104.

이제복・최상옥. 2018. “공공서비스 인공지능 ML 적용과 공공가치.” ≪정부학연구≫, 24(1):3-27.

이종한・황하・심우현・홍승헌. 2023. ≪스마트 규제를 위한 데이터플랫폼 구축 기본계획 수립 연구≫. 경제인문사회연구회.

장창기・성욱준. 2022. “인공지능 기반 공공서비스 정책수용 의도에 관한 연구: 개인의 인식과 디지털 리터러시 수준이 미치는 영향을 중심으로.” ≪정보화정책≫,

29(1):60-83.

정소윤. 2019. “인공지능 기술의 행정 활용에 관한 연구동향 및 쟁점 분석.” ≪한국지역정보화학회지≫, 22(4):175-207.

조인성. 2022. “행정 분야에서 인공지능(AI)의 적용 가능성-독일에서의 논의를 중심으로.” ≪법학연구≫, 25(1):77-104.

한명성. 2021. “정부의 인공지능(AI) 기반 서비스에 대한 국민의 사용 의향 분석: 공공가치와 확장된 기술수용모형을 중심으로.” ≪한국콘텐츠학회논문지≫, 21(8):388-402.

Amershi, Saleema, Cakmak, Maya, Knox, William B., & Kulesza, Todd. 2014. “Power to the people: The role of humans in interactive machine learning.” *AI Magazine*, 35(4):105-120.

Benington, John, & Moore, Mark. 2011. *Public value: Theory and practice*. New York: Palgrave Macmillan.

Braithwaite, Valerie. 2020. “Beyond the bubble that is Robodebt: How governments that lose integrity threaten democracy.” *Australian Journal of Social Issues*, 55:42-259.

Bryson, Joanna, & Winfield, Alan. 2017. “Standardizing ethical design for artificial intelligence and autonomous systems.” *Computer*, 50(5):116-119.

Cath, Corinne, Wachter, Sandra, Mittelstadt, Brent, Taddeo, Mariarosaria, & Floridi, Luciano. 2018. “Artificial Intelligence and the ‘Good Society’: the US, EU, and UK approach.” *Science and Engineering Ethics*, 24:505-528.

Commonwealth of Australia. 2023. *Report: Royal Commission into the Robodebt Scheme*.

Crawford, Kate. 2016. “Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics.” *Science, Technology, & Human Values*, 41(1):77-92.

Edwards, Lilian, & Veale, Michael. 2017. “Slave to the Algorithm? Why a ‘Right to an Explanation’ is Probably Not the Remedy You Are Looking For.” *Duke Law & Technology Review*, 16(1):18-84.

Fails, Jerry A., & Olsen, Dan R. 2003. “Interactive machine learning.” *Proceedings of the 8th international conference on Intelligent user*

interfaces. pp.39-45.

Gotterbarn, Don. 2010. "Thinking professionally: When soon after is way too late: the deception of 'opt-out'systems." *ACM SIGCSE Bulletin* 41(4):6-8.

Goodman, Bryce, & Flaxman, Seth. 2017. "European Union regulations on algorithmic decision-making and a "right to explanation"." *AI Magazine*, 38(3):50-57.

Henriques-Gomes, Luke. 2020. 'Robodebt: government admits it will be forced to refund $550m under botched scheme', The Guardian Australia, 27 March, https://www.theguardian.com/australia-news/2020/mar/27/ robodebt-government-admits-it-will-be-forced-to-refund-550m-under-botched-scheme (accessed 10 June 2024).

Holmberg, K., Dodig-Crnkovic, G., & Mueller, V. 2020. "Ethics in AI and Robotics: A Research Agenda." *AI & Society*, 35(4), 823-836.

Johnson, K. (2018). This Tool Could Protect Against AI Bias—If It Were Used. *Fast Company*.

Jørgensen, Torben, & Bozeman, Barry. 2007. "Public Values: An Inventory." *Administration & Society*, 39(3):354-381.

Monarch, Robert. 2021. *Human-in-the-loop Machine Learning: Active learning and annotation for human-centered AI*. Manning Publications.

Murphy, Kristina. 2019. "The robodebt horror was all about boosting the budget. That's the brutal truth," *The Guardian Australia*, 30 November, https://www.theguardian.com/australia-news/2019/nov/30/the-robodebt-horror-wasall-about-boosting-the-budget-thats-the-brutal-truth (accessed 10 June 2024).

O'Donovan, Darren. 2019. "Submission No. 15 to the Senate Community Affairs References Committee Inquiry into Centrelink's compliance program." https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Community_Affairs/Centrelinkcompliance/Submissions (accessed 18 June 2024).

Osborne, Stephen, Radnor, Zoe, & Nasi, Greta. 2013. "A new theory for public service management? Toward a (public) service-dominant approach." *American Review of Public Administration*, 43(2):135-158.

Settles, Burr. 2009. *Active Learning Literature Survey*. University of Wisconsin, Madison, 52(55-66), 11.

Simard, Patrice, Amershi, Saleema, Chickering, David, Pelton, Alicia, Ghorashi, Soroushi, Meek, Christopher, Ramos, Gonzalo, Suh, Jina, Verwey, Johan, Wang, Mo, & Wernsing, John. 2017. "Machine Teaching: A New Paradigm for Building Machine Learning Systems." arXiv preprint arXiv:1707.06742.

Wang, Jiangtao, Guo, Bin, & Chen, Liming. 2022. "Human-in-the-loop Machine Learning: A Macro-Micro Perspective." arXiv preprint arXiv:2202.10564.

Digital Transformation for Whom? Risks of Automated Welfare Services

Seung-Hun Hong & Ha Hwang

The rapid adoption of public services based on emerging technologies such as artificial intelligence is bringing both opportunities and risks. On the one hand, it makes public service processes more efficient and effective by enabling the collection and processing of vast amounts of data known to consume excessive resources and manpower in the past. It also enables the identification of blind spots, more scientific decision support, and customer-friendly service experiences. On the other hand, it raises challenges such as handling of private information, algorithmic bias, and cybersecurity. This paper focuses on risks that AI-based public services might pose to humans and discusses how human intervention is needed in the process of designing and introducing such services. Focusing on the Robodebt scandal in Australia, a representative case where automated welfare services adversely affect socially vulnerable welfare recipients, this paper discusses various risks inherent in AI-based public services and proposes triple-loop human intervention as a condition for AI system design.

※ Keywords: Digital transformation, Welfare services, Artificial intelligence, Robodebt, Tirple-loop human intervention