커뮤니케이션 이론
2023년 19권 2호 55~105쪽
https://doi.org/10.20879/ct.2023.19.2.055

공문을 보내면 네이버는 검색 알고리즘을 바꾸는가?
알고리즘 책무성 관련 증거 기반 논의*

윤호영 이화여자대학교 커뮤니케이션·미디어학부 조교수**

진보래 중부대학교 미디어커뮤니케이션학과 조교수***

본 연구는 알고리즘의 편향성 또는 공정성 문제와 관련하여, 알고리즘 운영 결과를 바탕으로 검증하는 책무성 기반 검증의 필요성을 제기한다. 근간에 사회적으로 주목받았던 알고리즘 검증 관련 이슈들을 살펴보면, 운영 기관이 알고리즘의 내용을 설명하는 투명성을 추구하면서 알고리즘이 특정한 결과를 유도하지 않는다는 방식의 해명이 주를 이루고 있다. 그리고 이 과정에서 투명성과 책무성을 실천하기 위해 노력했음을 역설하기도 한다. 그러나, 본 연구에서는 알고리즘의 내용이 특별한 설명 없이 조용히 변경되어 운영된 사례를 실제 데이터를 통해 살펴보면서 알고리즘 검증은 투명성 설명으로만은 부족하며 알고리즘 적용 후 나타나는 결과 및 알고리즘 변경에 대한 끊임없는 설명이 동반되는 책무성 기반 논의가 중요하다는 점을 논증했다. 최근 네이버는 한 시민단체의 문제제기 이후 선택적으로 이미지 검색 결과를 변경하였는데, 변경 내용이나 이유에 대해 아무런 설명을 제공하지 않고 있다. 이에 본 연구에서는 알고리즘 변경 전후의 이미지 검색 결과를 수집하여 알고리즘이 어떤 방식으로 변경되었는지 추정하였다. 분석 결과, 이미지 검색 알고리즘이 개선된 것이 아닌, 문제가 된 항목들의 일부만 '복합어 검색 결과로의 결과 교체', '유사어 검색 결과로의 대체', 그리고 '자동완성 검색어로의 대체'가 이루어진 것을 확인했다. 이 결과는 알고리즘 자체 투명성 기반 검증만으로는 알고리즘의 편향성 검증은 불충

* 이 연구는 아모레퍼시픽재단의 학술연구비 지원을 받아 수행되었음.

** hoyoungyoon@ewha.ac.kr

*** bjin23@joongbu.ac.kr, 교신저자

분하다는 점을 보여줌과 동시에 이와 관련된 사회적 논의의 필요성을 제기하고 있다. 본 연구는 향후 알고리즘 검증이 투명성 중심이 아닌 책무성을 기반으로 하여 알고리즘이 작동한 결과를 평가하고 모니터링하는 방식으로 이루어져야 할 필요성을 실제 데이터를 통해 확인했다는 데 의의가 있다.

KEYWORDS 알고리즘, 투명성, 책임성, 책무성, 네이버, 영향 평가

1. 서론

인터넷 트렌드 데이터의 분석 결과에 따르면, 2022년 전체 기간 평균 우리나라 검색 엔진 시장 점유율은 네이버가 61.2%, 뒤이어 구글이 28.6%를 차지하고 있다(Internet Trend, 2022). 우리가 매일 경험하는 현실은 네이버, 구글, 인스타그램, 유튜브 같은 플랫폼 기업이 제공하는 뉴스, 친구 소식, 이미지, 영상, 그리고 검색 결과에 의존하고 있다. 플랫폼 서비스가 과거 언론의 역할이라 여겨졌던 문지기(gatekeeper) 기능을 수행해 온 지는 오래되었고. 검색 엔진은 이용자를 만족시키기 위한 "편집상의 선택(editorial choices)"을 하는 미디어이기도 하다(Goldman, 2006, p. 189).

2021년 6월 국회에서 정보통신망법 일부 개정안이 발의되었다. 해당 개정안은 알고리즘 투명성위원회를 방송통신위원회 산하에 설치하고, 알고리즘 서비스 이용자가 서비스 제공자에게 알고리즘 설명을 요구할 수 있도록 하며, 제공자는 영업비밀을 이유로 이를 거부할 수 없도록 하는 등의 내용을 담고 있어, 알고리즘 투명화법이라 불리고 있다. 같은 해 11월에는 "알고리즘 및 인공지능에 대한 법률안"이 발의되었는데, 이 법안은 "국민의 생명, 신체의 안전 및 기본권의 보호에 중대한 영향을 미치는 인공지능"을 고위험 인공지능이라 정의하고 그 범위를 구체화한 후 인공지능을 개발할 때 준수해야 할 기본 원칙을 명시했다. 방송통신위원회는 새 정부 출범 직후, "포털 뉴스 신뢰성·투명성 제고를 위한 협의체"를 구성하여 포털 중심의 뉴스 서비스 생태계의 투명성과 공정성을 강화하는 방안을 마련하겠다고 발표했다(유진상, 2022). 이러한 빅데이터, 알고리즘, 인공지능 관련 규제에 관한 논의는 우리나라에 국한되지 않는 세계적인 추세다. 미국 뉴욕시에서는 2017년 알고리즘 책무성 법률안(Algorithmic Accountability Bill)이라는 이름으로 알고리즘 관련 법률이 세계 최초로 발의되었고(Bernard, 2017), 상원에서도 2019년에

알고리즘 책무성 법안(Algorithmic Accountability Act)을 발의했다가, 2022년 상하원 공동으로 개정 버전이 다시 상정되었다(Khalid, 2022). 미국 백악관에서는 시민의 권리를 보호하고 인공지능의 책임을 묻기 위한 인공지능 권리장전(AI Bill of Rights)을 발표하였다(Heikkilä, 2022). 유럽 연합은 2018년에 발효된 일반 정보보호 규정(General Data Protection Regulations)을 집행해 온 경험을 바탕으로, 2021년 4월 인공지능 법안(AI Act)의 초안을 공표하였고 이후 개별 조항들이 제정되고 있다(정소영, 2022). 법률안 발의와 더불어 국내에서는 알고리즘에 대한 검증 작업도 진행중이다. 네이버 뉴스 알고리즘, 카카오T의 택시 배차 알고리즘, 배달의 민족의 배차 및 거리 측정 알고리즘 등에 대한 문제가 제기되고 이어서 검증도 이루어졌다. 한편 네이버의 쇼핑 검색 알고리즘에 대한 공정위의 처분과 MBC의 네이버 뉴스 알고리즘에 관한 보도 내용은 소송 중에 있다. 알고리즘을 개발한 당사자가 스스로 검증하는 자율 검증의 경우, 대부분 자체 위원회를 구성하여 편향성 시비와 같은 제기된 문제를 조사하지만 대체로 알고리즘 자체는 편향적으로 설계되지 않았다고 결론짓는다(김국배, 2022; 최창원, 2022).

하지만, 알고리즘 자체에 편향이 내재하고 있는지와 알고리즘이 구현되고 작동한 결과로 편향이 발생하는지는 다른 문제이고 서로 구분되어야 한다. 예를 들어, 구글 포토 앱이 흑인 커플의 얼굴을 고릴라로 인식하여 구글이 사과한 적이 있는데(Dougherty, 2015), 구글이 흑인을 고릴라로 인식하도록 이미지 인식 알고리즘을 구성했다고 보기는 어려우나 그러한 결과가 발생한 것이다. 또한, 구글 광고가 흑인 이름의 사용자에게는 법률 상담 광고를 더 많이 내보내고(Sweeney, 2013), 여성보다 남성에게 임금이 높은 전문직 취업 광고를 더 많이 보여준다는 것이 밝혀지기도 했다(Datta, Tschantz, & Datta, 2015). 알고리즘의 편향성과 관련된 논쟁은 미국 법원에서 활용하는 COMPAS라는 알고리즘 사례에서 매우 극심했다. 백인 대비 흑인에게 더 가혹한 판결을 냈다는 프로

퍼블리카의 보도와 뒤이은 노스포인트社의 반박 그리고 대법원 항소 등 일련의 과정에서 알고리즘이 가지는 공정성을 어떻게 판단하고 평가할 것인지에 대한 격론이 벌어졌다(오요한·홍성욱, 2018; Angwin, Larson, Mattu, & Kirchner, 2016; Berk, Heidari, Jabbari, Kearns, & Roth, 2021; Dieterich, Mendoza, & Brennan, 2016; Dressel & Farid, 2018; Liu, Lin, & Chen, 2019).

본 논문은 이와 같은 사례들처럼 알고리즘이 어떻게 작동되고 있는지에 대한 주로 투명성에 기반한 검증이 반드시 해당 알고리즘이 운영되는 과정에서 나타나는—때로 조작이라고까지 부를 수 있는 불투명한—알고리즘 변경 및 운영에 관한 실질적인 검증이 되지 않는다는 점을 네이버 이미지 검색 사례를 통해 보여주고자 한다. 알고리즘에 대한 평가와 검증은 알고리즘이 작동하는 원리에 대한 조사나 투명성에서 종료되는 것이 아니라, 알고리즘이 보여주는 결과에 기반한 책무성을 기준으로 이루어져야 하며, 알고리즘 운영과 관련된 모니터링이 포함되어야 함을 제시하고자 한다. 이는 알고리즘 투명성 확보가 알고리즘을 둘러싼 책임의 문제를 해소할 수 없다는 최근의 논의와 궤를 같이하는 것이면서(Edwards & Veale, 2017; Johnson, 2022), 동시에 앞서 서술한 단순히 투명성에만 기반한 정책 방향 그리고 알고리즘을 둘러싼 사회적 논란이 한계가 있음을 지적하고자 한다.

구체적으로 본 연구는 국내 대표적인 플랫폼 기업 네이버가 이미지 검색 결과를 수정하여 검색 결과를 인위적으로 바꾼 사례들을 수집하였다. 이 수정은 한 시민단체가 특정한 검색 결과들에 대한 문제를 제기한 후에 이루어졌다. 그러나 네이버는 알고리즘의 변경과 관련해 아무런 설명을 제공하지 않았으며 임시방편적인 방식으로 검색 결과를 바꾼 것으로 나타났다. 이 사례는 알고리즘의 변경 및 적용과 관련된 학술적, 사회-정책적인 측면에서 다양한 질문을 제기하고 있다. 또한 알고리즘의 투명성을 강조하는 것이 공정하거나 설명가능한 알고리즘으로 이어지지 않고,

오히려 불투명하고 불완전할 수밖에 없는 플랫폼 기업의 사회적·윤리적 책임을 덜어줄 수 있음을 보여준다고 볼 수 있다. 이와 관련하여, 저자들이 아는 한 플랫폼 회사 내부의 데이터를 직접 얻어 분석한 후 공개적으로 발표한 연구가 없었고, 수사권과 같이 특별한 권리가 없는 상태에서 국내에서 알고리즘 변경과 관련된 내용을 실제 근거 데이터를 수집하여 검증하고자 하는 시도는 처음이기에 본 연구가 새로운 논의를 촉발하는 계기가 될 수 있을 것으로 기대한다.

본격적인 논의에서 앞서 몇 가지 사항을 미리 언급하고자 한다. 우선, 이 글에서는 알고리즘 변경과 알고리즘 운영상의 변경을 구분하지 않는다. 알고리즘 운영의 변경은 알고리즘을 언제 어디에 어떻게 적용할 것인가에 문제인데 이는 결과적으로 알고리즘이 변경된 것과 동일한 효과를 가지기 때문이다. 예를 들어, 검색 알고리즘 자체를 수정하지 않고 특정한 검색어에 대해서만 알고리즘이 내놓는 결과를 인위적으로 바꾼다면, 알고리즘이 본래 가지고 있는 절차를 변경하는 것임과 동시에 알고리즘이 운영되는 방식을 변경한 것이기도 하다. 따라서 이 글에서 정의하는 알고리즘 변경은 알고리즘의 운영 변경을 포함한다. 두 번째로, 알고리즘과 인공지능(AI) 시스템을 모두 알고리즘으로 통칭하고 알고리즘 윤리와 인공지능 윤리 역시 모두 알고리즘 윤리로 통합하여 이해하고자 한다. 인공지능이나 알고리즘 모두 의사결정을 위한 시스템이라는 공통적 성격을 가지고 있으며 기존 문헌에서도 알고리즘 윤리와 인공지능 윤리를 동일한 것으로 보고 메타 분석이 진행된 바 있다(예, Jobin, Ienca, & Vayena, 2019).

논문은 크게 세 부분으로 구성되었다. 먼저 알고리즘 편향을 둘러싼 논점 그리고 알고리즘 검증 준칙과 관련된 투명성과 책무성을 다루면서, 본 연구가 검증하는 내용 및 주장하는 바에 대한 최근 논의를 점검한다. 이어서 네이버 사례에 대한 분석이 가지는 의의를 서술하면서 분석 방법 및 분석 결과를 제시한다. 마지막으로 본 분석 결과가 보여주는 함의를 논하면서 마무리하고자 한다.

2. 알고리즘 검증과 관련된 기존 논의

1) 알고리즘 편향: 규명을 둘러싼 논점

알고리즘의 잠재적 영향을 실제로 측정하고 결정하는 것은 매우 어려운 작업이다. 미텔슈타트와 동료들(Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016)은 그 이유를 우선 알고리즘이 가진 성격에서 찾았다. 무엇보다 알고리즘은 학습에 따른 결과가 완전히 확실한 성격을 가지는 것이 아니라 확률에 근거한 불확실성을 가지기에 비결정적이고, 그에 따라 알고리즘에 제약을 가하기 위한 행동의 정당성이 약해진다는 것이다. 이는 알고리즘에 의한 데이터 처리와 결과를 사람이 이해할 수 없어 불투명해지는 결과로 이어지는데(Tsamados et al., 2021), 통상 블랙박스(Black Box)라 부르는 알고리즘에 대한 이해가능성과 해석가능성의 부재는 결국 알고리즘을 어떻게 제어, 감시 및 교정해야 하는지 모호하게 만든다. 이른바 무작위 확률에 따른 비결정성이 가져오는 해석불가능성이다(Vedder & Naudts, 2017).

미텔슈타트와 연구진(Mittelstadt et al., 2016)은 알고리즘이 작동하기 위해 입력되는 데이터 또는 학습 데이터라 불리는 데이터의 문제와 알고리즘이 작동하여 나타나는 결과의 문제 모두가 알고리즘 윤리와 결부되어 있다고 하였다. 대표적으로 잘못된 데이터를 투입하면 잘못된 결과가 나오는(Garbage in, Garbage out) 경우를 예로 든다. 여기서 중요한 것은 어떤 데이터가 좋은 데이터인지 최초로 판단하는 것부터 시작하여, 데이터를 알고리즘에 투입하여 나오는 결과를 통해 사후에 데이터가 좋은 데이터인지 아닌지 판단하는 것 모두가 이 문제와 관련된다는 것이다. 사후에 적용되는 데이터 판단과 관련하여 아마존 채용 알고리즘을 이야기할 수 있다. 아마존은 직원들의 데이터를 바탕으로 구직자의 이력서를 평가하여 채용에 적용하는 알고리즘을 만들었는데, 알고리즘 적용 결과 여성이 체계적으로 배제되는 결과가 발생했고 해당 프로젝트를

중단했다(Dastin, 2018). 실제 데이터를 보면 회사에 남성이 많은 것 자체는 사실이기 때문에, 데이터를 그대로 이용한 점은 가치 판단이 없는 중립성을 띠고 있다고 볼 수 있으며 또한 이를 알고리즘에 적용한 것도 합당하다고 판단할 수 있다. 하지만 그 결과 알고리즘의 효율적인 학습이 기존의 차별성을 더욱 강화시키는 방향으로 나타나게 된다. 결국 데이터 문제는 알고리즘의 장기적 영향과 비가역성의 문제로 이어진다(Mittelstadt et al., 2016; Tsamados et al., 2021). 완벽하게 보이는 알고리즘이라도 특정 집단에게는 불공정하게 작용할 수 있는데, 이러한 편향이 당장은 큰 문제를 일으키지 않아 무해한 듯 하지만 장기적인 축적 결과가 어떤지는 예측하기 어렵다. 또한 알고리즘과 관련된 윤리적 문제가 누구에게 있는지 알기 어려우며, 개인은 전체 맥락에 대한 이해도 어렵다(Mittelstadt et al., 2016; Tsamados et al., 2021). 한 번 적용되면 이미 그 영향력이 발휘되어 되돌리기 어려운 변형적 효과(transformative effect)를 가진 상태가 된다. 예를 들어, 핀테크 기술에 의해 최초 대출을 갚을 수 있는 능력이 없다고 낙인찍히는 경우 아예 대출을 평생 받을 수 없게 될 수 있다. 종합하면, 알고리즘을 적용하는 과정에서 나타나는 편향은 어떤 때는 알고리즘의 데이터 처리 때문이기도 하고 어떤 때는 (준)자동화된 결정에 있기도 하고 어떤 때는 윤리적으로 중립적이어도 나타나기도 하는 등 명확하게 책임 소재를 찾기 어렵다.

아마존 사례에서 보듯, 알고리즘의 편향을 논의하는 과정에서 알고리즘이 공정한가에 대한 판단은 알고리즘이 어떠한 과정을 거쳐서 결과를 내놓는가에 대한 판단과 알고리즘이 내놓는 결과에 대한 책무는 어떻게 규정할 것인가에 대한 문제로도 이어진다. 알고리즘 윤리에 관한 문헌에 따르면, 알고리즘의 윤리와 관련된 대부분의 논의가 투명성, 공정성, 책임성의 3가지 원칙으로 집중된다(Jobin et al., 2019). 사실 이들 3가지 원칙은 불가분의 관계로 서로 맞물려 있다. 예를 들어, 범죄자의 재범 확률을 판단하여 양형에 반영함으로써 논쟁의 대상이 되었던

COMPAS 알고리즘을 생각해 보면, 예측이 공정한가를 판단하기 위해서는 예측의 공정성을 어디에 둘 것인가 기준을 세우고, 예측의 결과가 그러한 기준에 부합하는지 판단해야 한다(오요한·홍성욱, 2018; Berk et al., 2021).

그러나, 예측의 공정성 기준이 이해 당사자들마다 다르다는 점이 공정성 판단을 어렵게 한다. 노스포인트社에게 공정성이란 각 인종 집단 내에서 재범 예측을 얼마나 정확하게 하는가를 의미하는 것이었다면, 프로퍼블리카는 재범자를 안전한 사람이라 잘못 분류한 비율은 백인이 높고, 재범자로 판정했으나 재범자가 되지 않은 사람의 비율은 흑인이 높은 오류의 편향성이 공정성을 판단하는 기준이었다. 공정성의 기준은 차치하더라도, 알고리즘이 공정하다는 점을 증명하기 위해서는 알고리즘이 어떻게 작동하고 있는지 설명할 수 있어야 하며 그에 따른 알고리즘의 투명성을 확보해야 할 필요성이 생긴다. 만약 알고리즘의 투명성이 확보된다면 공정성에 기반한 책무성에 대한 판단도 내릴 수 있다. 설명된 알고리즘 논리에 기반하여 공정성을 확인할 수 있기 때문이다. 그러나, 알고리즘의 투명성이 보장되지 않은 상태라면 알고리즘이 작동하여 보여주는 결과를 통해 공정성을 판단하고, 문제제기를 통해 알고리즘의 결정과 관련된 책무가 어디에 있는지 또는 누구에게 있는지 판단해야 한다. 즉, 알고리즘의 결과에 대한 평가와 판단은 데이터 처리 과정에 중점을 두는 투명성과 결과에 상당 부분 무게를 두는 책무성 두 가지 접근이 필수적으로 요구된다.

2) 알고리즘 평가(Assessment): 투명성과 책무성의 두 가지 접근법

(1) 알고리즘의 투명성 기반 접근법

투명성은 기본적으로 특정한 사항과 연관된 정보를 모두 공개하는 것을 의미한다. 이러한 투명성에 대한 요구는 전통적으로 정부나 기업과 같은

기관을 대상으로 했다(Meijer, 2014). 이때 투명성은 정보의 비대칭성을 극복하고 관련자들에게 모든 정보를 제공함으로써 합리적인 의사결정을 내릴 수 있도록 하기 위한 것이다(Forssbaeck & Oxelheim, 2014). 이러한 기본적인 투명성에 대한 이해가 책무성과 연관되면서 기관이나 조직과의 관계로 확장될 경우 투명성은 "특정한 행위자가 다른 행위자의 업무나 행위를 감시할 수 있도록 하는 정보의 이용가능성(availability of information about an actor allowing other actors to monitor the workings or performance of this actor)"으로 정의된다(Meijer, 2014, p. 512). 여기서 투명성은 모든 것이 분명하고 명확하게 종결된 상태가 아니라, 관찰 가능하고 통제 가능함을 인식하는 것을 말한다(Ananny & Crawford, 2018).[1] 이는 다시 어항 투명성(fishbowl transparency)과 합리적 투명성(reasoned transparency)로 구분되기도 하는데. 어항 투명성은 어항을 보듯 모든 것을 보는, 마치 공중이 정부의 운영을 문서와 청문회 및 정보 공개 등을 통해 살펴보듯이 참여하는 접근권 중심의 관점이며, 합리적 투명성은 제공된 정보의 유용성을 강조하는 관점이다(Coglianese & Lehr, 2019). 다시 말해서, 특정한 정보를 '공개'하는 투명성의 요건과 공개된 정보가 실질적으로 유의미한 정보인지 '유의미성'으로 투명성을 판단하는 셈이기에 투명성은 '공개 여부'가 기본 바탕이 된다.

알고리즘 투명성은 알고리즘 윤리와 관련하여 가장 많이 언급되는 원칙으로서 데이터 활용, 인간-AI 상호작용, 자동화된 결정, AI 시스템 응용 또는 데이터 사용 목적 등 매우 다양한 맥락과 관련되어 있다. 알고리즘 투명성은 편향된 의사결정으로 인한 피해를 최소화하고 동시에 AI

---

1) 투명성을 기관이나 조직과 맺는 관계가 아닌 그 자체를 하나의 가치로 보는 관점도 존재한다. 이 관점에 따르면, 투명성이란 규범적인 관념으로서 공적인 행위자의 행동을 평가하는 기준을 정해놓고 모든 행위를 공개하는 것으로, 특별히 투명하게 공개할 대상을 사전에 정해놓는 것은 아니다(Felzmann, Fosch-Villaronga, Lutz, & Tamo-Larrieux, 2020).

성능 향상과 법적인 규제 및 신뢰 향상을 위해 추구되며(Mittelstadt, 2019), 설명가능성(explanability)과 설명요구권(the right to explain)으로 구성된다(황용석·정재선·황현정·김형준, 2021). 여기서 공개의 문제는 알고리즘이 어떻게 작동하는지 그 원리를 사람들이 납득할 수 있도록 설명하는 것에 집중된다. 이 가운데 설명가능성은 주로 민간분야에서 내놓는 기술적인 해결책으로 여겨진다. 알고리즘이 어떻게 작동하는지에 대한 기술적인 설명을 함으로써 그리고 알고리즘이 어떻게 작동하는지 설명가능한 모델을 만들어 비교함으로써, 알고리즘이 내놓는 결과에 대한 인과관계를 확인하는 등, 이렇게 투명한 절차를 거쳐서 나오는 결과를 통해 알고리즘의 공정함과 신뢰성을 확보하고자 하는 것이다(Arrieta et al., 2020). 유럽 연합의 일반 정보보호 규정이 이러한 접근을 취하고 있는데, 이러한 접근은 투명성이 책무성을 담보할 수 있다는 가정을 담고 있다(Edwards & Veale, 2017). 즉, 투명성을 통해 알고리즘을 평가하는 것은 알고리즘의 작동 원리를 보여줌으로써, 알고리즘을 적용하면 어떠한 편향이나 차별을 만들어 낼 수 있는지 사전에 검증하는 방식을 취한다.

그러나, AI 관련 윤리 연구를 한 아나니와 크로포드(Ananny & Crawford, 2018)는 투명성을 보는 것(seeing)과 아는 것(knowing)을 동일시하는 '이상(ideal)'에 불과한 것이라 보면서, 여러 가지 이유로 투명성이 문제 해결책이 되기는 어렵다고 비판한다. 무엇보다 투명성 그 자체는 정보 비대칭으로 인해 발생하는 권력의 불균형과 같은 문제를 해결하는 데 도움이 되지 않는다는 것이다. 또한 투명성이 생긴다고 해서 데이터 원천이 가진 문제까지 추적해 나가는 것은 많은 경우 불가능하다. 예를 들어, 뉴스 알고리즘이 정치적 편향성에 따라 뉴스를 추천하는 것은 아니지만 뉴스 생태계의 불균형을 그대로 반영하는 경우, 해당 알고리즘의 논리가 밝혀진다고 해서 그 알고리즘에 의해 뉴스 생태계의 불균형이 더욱 확산하는 문제가 해결되는 것은 아니다. 또한 투명성이 오히려 해악

이 되는 경우도 존재한다. 투명하게 모든 것을 공개하는 것에만 집중하다 보면 취약한 개인이나 집단을 위험에 빠트리게 될 수 있고, 알고리즘이 남용될 수 있다. 예를 들어, 개인의 민감한 정보가 강제로 노출되는 것과 같이 프라이버시를 침해하거나, 공개된 알고리즘이 악용되어 스팸이나 허위 정보가 대량 유포될 수 있다. 실제로 검색 알고리즘이나 포털 사이트의 콘텐츠 배열 알고리즘이 전부 공개된다면 이를 남용하고자 하는 시도가 빗발치게 될 것이다. 마지막으로, 알고리즘을 공개하는 과정에서 의도적으로 보여주고 싶은 것만 보여주는 것도 가능하다. 투명하다는 것이 반드시 신뢰를 가져오는 것은 아니라는 것이다. 게다가 알고리즘이 가진 복잡성은 문제를 더욱 배가시킨다. 특정 시점에 공개된 알고리즘은, 종종 확률적 결정이라는 성격과 결합하여, 이후 시점에도 같은 결과를 낸다는 보장이 없다(Ananny & Crawford, 2018).

최근에는 기술을 통해 설명 가능성을 확보하여 알고리즘을 평가하고 그에 따른 책무성을 이행하려는 움직임이 있기도 한데(Loi, Ferrario, & Viganò, 2020; Tan, Caruana, Hooker, & Lou., 2018), 문제는 알고리즘 평가와 관련된 책무성은 기술의 문제가 아니라 사회적인 문제라는 점이다. 비근한 예로, 이미지 인식 알고리즘이 원숭이를 고릴라로 잘못 인식하였다면 이는 기술적 문제라고 단순하게 취급될 수 있지만, 서론에서 언급했던 흑인을 고릴라로 인식하는 것은 기술상의 오류를 넘어서 사회정치적 문제가 된다. 이는 구글이 바로 공식적으로 사과했다는 점이 말해준다. 이러한 점 때문에, 실제 내부 작동 원리를 설명하는 것이 아닌 그것이 작동한 결과, 즉 입력값에 따라 산출된 결과를 보는 것이 현실적으로 합당한 책무성을 확보하는 방법이라 지적된다(Kroll et al., 2017). 현실 세계의 법정에서도 유사한 논리가 적용되는데, 사건의 세부적인 모든 과정을 알지 못해도, 목격한 결과와 부분적인 증거 및 상황으로 판결이 이루어지며, 투명성이 아닌 전문성이 보장되는 자격을 가진 사람들이 평가에 기여한다. 이 맥락에서 투명성의 효용은 크지 않다.

(2) 알고리즘의 책무성 기반 접근법

알고리즘 작동원리에 대한 설명을 중심으로 한 투명성과 달리, 책임과 책무는 자유 의지를 가지지 않은 행위 주체, 즉 인공지능에게 적용할 수 있는가라는 의문이 제기될 수 있다. 전통적으로 도덕적 책임에서와 같은 책임은 행위 주체가 자유 의지를 가지고 있음을 전제한다. 따라서 알고리즘 같은 시스템의 행동에 대해서 책임을 묻기는 어려우나, 책무는 의무의 묶음으로서 인간이 아닌 행위자에게 적용할 수 있다(이중원, 2019). 이중원에 따르면, 자율주행차에 의한 사고 발생 시 이에 대한 설명 요청이 있는 경우 자율주행차의 시스템을 기반으로 응답해야 할 책무가 생긴다. 그리고 국가기관이나 사회 제도 차원에서 인공지능 시스템이 법적 책무도 지게 된다.

이러한 책무성과 관련하여 가장 광범위하게 쓰이고 있는 보벤스(Bovens, 2007)의 정의에 따르면 책무성이란 자신의 행동을 설명하고 정당화해야 하는 의무가 있는 행위자(an actor)와 질문을 제기하고 판단을 내릴 수 있는 청중(a forum) 간의 사회적 관계이다. 여기서 말하는 행위에 대한 설명의 핵심은 특정 행위에 대한 '정당성'을 확보하기 위한 근거를 제시하는 것이고, 해당 근거를 바탕으로 행위에 대한 판단을 내리는 과정 속에 상호 합의하는 결과를 얻어내는 것이 책무성이다. 이 과정을 보면, 투명성 자체로 책무성이 완성되는 것이 아니라, 투명성을 기반으로 판단을 위한 문제제기와 논의를 거쳐서 결정을 내리는 일종의 합의나 협의 과정 그리고 이를 받아들이는 전 과정이 책무성을 구성한다. 알고리즘 책무성에 관한 미국 상원에 최근 제출된 법률안 역시 기업들이 자동화된 시스템을 사용하거나 판매할 때 자동화 시스템의 영향을 평가하도록 요구하였는데, 언제 어떻게 자동화된 시스템이 사용되는지에 관해 소비자에게 투명하게 알리고, 중요한 결정이 자동화되는 것에 대한 정보를 제공하여 소비자들이 판단할 수 있도록 해야 한다(Wyden, 2022). 즉, 영향 평가의 성격을 지닌 내용을 공개하고 이를 기반으로 판단할 수

있도록 하는 방식이 책무성의 중요한 요소이다.

그러나 보벤스(2007)가 정의하는 식의 책무성이 알고리즘과 관련해서는 그대로 적용되기 어려운 때가 있다. 무엇보다 행위자와 청중 간에 설명할 의무가 있다는 믿음에 기반한 규범이 공유되어야 하는데, 인공지능이나 알고리즘의 경우 행위자가 동의하지 않는 경우가 발생한다(Johnson, 2021). 상업 알고리즘을 운영하는 회사는 영업비밀을 이유로 알고리즘을 공개하지 않을 수 있으며 알고리즘에 의한 결과에 대한 설명도 사람이 결론을 내리는 것이 아니라며 거부하곤 한다. 존슨은 보벤스의 정의가 가진 한계를 지적하면서 공식적인 법이나 규제 또는 정책에 서술되지 않은 결과는 합의에 도달하기 어렵기 때문에, 책무성은 결과적으로 행위를 제한하는 실천의 영역에 속하며, 어떤 실천을 수행하는가에 의해서 정의되는 것이라고 주장한다. 책무성 주체의 문제 역시 알고리즘은 지속적으로 학습되고 많은 사람들이 관여하기 때문에, 좁게는 알고리즘 설계자와 사용자가 책임을 종종 나눠가지게 되면서 책임성의 주체가 형해화된다고 지적한다. 알고리즘을 만든 주체인 행위자에게 설명하고 정당화해야 하는 의무가 주어진다는 전제 자체가 적용되지 않을 수 있는 것이다(Johnson, 2021). 다시 말해, 현재 알고리즘 책무성의 가장 큰 문제점은 알고리즘의 설계와 사용에 있어서 규범이 부재하는 것이다. 이와 관련하여 보벤스의 책무성에 대한 논의에 근거하여 알고리즘 책무성에 관한 메타 분석을 시도한 비에링아(Wieringa, 2020)는 알고리즘의 책무성을 알고리즘의 존재 이유, 개발의 맥락, 그리고 효과를 망라하는 개념으로 보아야 하고 일회적이거나 이분법적이 아닌 알고리즘의 기획과 개발, 운영, 사후 평가 전 과정에 걸쳐 지속되어야 하는 작업이어야 한다고 말한다. 투명성으로 이해하자면, 설명이 알고리즘 내부 로직에 한정되지 않고 광범위하게 확장되어, 알고리즘 개발과 운영, 보수 전 과정에 있어서 보장되어야 하는 것이다. 이렇게 포괄적인 측면에서의 투명성은 책무성을 보장할 수 있다는 뜻이다.

실제로 AI 알고리즘의 책무성과 관련된 논의는 첫 번째 단계를 지나 두 번째의 새로운 단계(Second Wave)로 진행되고 있는데, 그 과정에서 알고리즘의 원리와 논리가 어떠한지는 부차적인 문제로 취급된다(Pasquale, 2019). 책무성의 첫 단계(First Wave)에서는 AI 알고리즘 시스템이 어떤 시스템이어야 하는지, 어떤 문제가 해결되어야 하고, 누가 만들어야 하는지, 누가 결정하는지 등에 관한 물음으로부터 시작하여, 기업 외부에서 접근 가능한 책무성 구조를 만들어야 한다는 논의가 주류였다. 이 단계에서는 인간의 삶에 통합되는 AI 알고리즘은 검증되어야 하고 책임이 분명해야 할 뿐만 아니라 공중의 이해를 대변해야 한다고 믿고, 알고리즘의 공정성 및 불편부당함을 검증하고 교정하는 작업을 중요하게 여긴다. 예를 들어, 소수자의 얼굴인식이 안되는 경우라면 이들을 데이터로 포함시켜 공정한 알고리즘을 만들어야 한다는 것이다.

책무성 논의 두 번째 단계에서는 알고리즘의 원리가 아니라 알고리즘을 적용하는 과정에서 나타나는 문제에 보다 집중한다. 알고리즘이 보여주는 문제들을 검증하고 해당 문제들을 수정하기 위해 어떠한 작업이 필요한지 논의하는 것이다. 구성적 기술 영향 평가(Schot & Rip, 1997) 원리와 유사한데, 단순히 만들어진 알고리즘을 평가하는 것이 아니라 해당 알고리즘 자체에 대한 문제제기를 통해 논의를 발전시키고자 하는 의도를 가진다. 예를 들어, 정신건강과 관련된 앱을 개발할 때, 앱이 다양한 커뮤니티나 소수자들을 모두 포용할 수 있는 방식으로 만들어져야 한다고 보고 이를 기준으로 평가하는 것이 1단계라면, 고비용 전문가 중심의 작업이 저비용의 소프트웨어로 대체되면서 미숙한 알고리즘이 가져올 혼란에 대해 평가하는 것이 2번째 단계이다(Pasquale, 2019). 이 방식은 알고리즘에 의해 나타나는 소외된 상황에 대한 문제제기를 가능하게 한다. 예를 들어, 아마존 채용 시스템에서 여성이라고 보이는 흔적이나 특질이 발견되면 채용에서 불이익을 받는다는 사실은 채용 지원을 하는 개인 지원자 입장에서는 전혀 파악할 수 없는 내용이기 때문에

차별을 받고 있는지 알 수도 없고, 사회 전체적으로 잘 드러나지도 않는다. 즉, 투명하게 알고리즘 작동원리를 설명한다고 해서 알고리즘이 보여주는 모든 편향을 사전에 알 수 없다. 그리고 알고리즘의 적용 결과 어떠한 경우에 편향이 나타나는지 관찰하기 어렵고, 실제로 그러한 사례가 존재한다고 하여도 흔히 이러한 편향은 개개인의 수준에서 경험되기 때문에, 집단적으로 논의되고 토론될 정도로 사회적인 증거가 드러나 쌓이지 않는 이상 모르고 넘어갈 가능성이 크다. 편향된 알고리즘에 의해 평가받았는데, 이러한 평가가 편향되거나 차별적인 것이라 여기기보다는 알고리즘이 정한 기준에 본인이 미치지 못했다고 판단하는 것이다. 예를 들어, 동일 노동에 대해 남성과 여성의 임금 차이를 두고 있음에도 서로 연봉을 공개하면서 이야기하지 않기 때문에 그러한 차별이 있는지조차 모르는 경우와 유사하다. 다시 말해, 알고리즘의 원리가 투명하다는 사실이 결과에서의 공정성과 정당성을 보장하지 않기 때문에 알고리즘을 적용할 때 나타날 수 있는 문제점들을 모두 검토하여 책무성을 수행하고자 하는 시각이다. 알고리즘의 사회적 책임과 관련된 논의를 종합한 연구에서는 알고리즘의 공정성과 포용성을 강조하면서, 알고리즘에 적용되는 책무성이 알고리즘 개발 및 적용 그리고 적용 이후 나타난 결과에 이르는 전 과정에서 모두 요구된다고 정리한다(Cheng, Varshney, & Liu, 2021).

지금까지 살펴본 알고리즘 책무성 논의는 알고리즘의 책무성이 알고리즘을 설명하는 일회적인 이벤트로 끝나야 하는 것이 아니라, 알고리즘이 적용되는 전 과정—소프트웨어 개발로 생각하면 소프트웨어 생애주기 전 과정—에 걸쳐서 끊임없이 검증되고 평가되어야 한다는 논지로 요약된다. 그리고 평가의 내용과 기준에 관한 사회적 합의를 위해서는 여러 이해 당사자 간 지속적인 논의와 토론이 필요하다고 본다. 이 과정에서 알고리즘 책무성에 관한 담론과 규범도 만들어가야 하는 것으로 설정된다. 실제로 유럽 연합의 알고리즘 관련 규제는 책무성을 AI가 달성하고자 했던 목표나 기능을 달성하지 못하거나 전례 없는 결과가 발생하는 것

에 대해 의도 여부와 관계없이 개발사와 제조자가 모두 책임을 져야 한다고 서술하면서, 이러한 책임은 단번에 끝나는 의무가 아니라 지속적으로 이루어져야 하는 알고리즘의 조정, 변화, 채택의 반복적인 과정이라 본다(European Union, 2021).

3. 알고리즘 검증 사례 연구: 알고리즘 검증에 관한 문제제기 및 연구 방법론

1) 문제제기

지금까지 우리 사회에서는 앞서 논의한 바와 같은 알고리즘에 대한 책무성 기반 검증이 잘 이루어지지 않고 있다. 본 연구는 바로 이 점에 착목하여 책무성 기반 검증을 시도해보고자 한다. 특히, 현재 우리 사회에서 알고리즘 특히 인공지능에 적용된 알고리즘과 관련된 논의는 기술 원리를 설명하는 투명성을 기반으로 한 기술 중심주의로 전개되는 경향이 있고, 사전에 알 수 없는 결과를 개발 단계에서 모두 알 것이라 가정하는 영향 평가를 응용하기에, 본 연구의 증거 기반 검증을 통한 문제제기는 현재의 논의를 더욱 발전시키는 과학적 반증(Popper, 2002)으로서 의미가 있을 것으로 본다.

실제로, 정부가 제시한 인공지능 실현 전략안은 '신뢰'를 기치로 하고 있다. 여기서 신뢰란 '설명 가능'한 기술 개발을 통해 '공정'을 강화하는 기술을 개발한다는 것을 의미한다(과학기술정보통신부, 2021). 국내 포털기업에 대한 알고리즘 검증에서도 투명성을 중심으로 설명한다. 예를 들어 뉴스 추천 알고리즘에 대한 문제제기에 대해 네이버는 "특정 성향에 유리하게 추천하는 것은 기술적으로 불가능"하며 "공정성 문제는 알고리즘 자체보다는 생산자와 사용자의 상호작용 속에서 발생한다"고 말한다(채새롬, 2021). 또한 네이버는 자체 AI 준칙을 소개하면서 "일상에서

AI의 관여가 있는 경우 사용자에게 그에 대한 합리적인 설명을 하기 위한 책무를 다하겠습니다" 그리고 "투명성(Transparency) 실현의 방법으로 설명 책무를 명시하였다"고 기술하고 있다(Naver, 2021). 이와 같은 설명에 기반한다면 기술적으로 알고리즘을 특정하게 바꾸는 것은 불가능한 것인지 그리고 변경 시 합리적인 설명이 제공되고 있는 것인지 살펴보는 것이 책무성 검증의 방법이 된다.

당연하게도 책무성 기반 검증에서 투명성이 불필요하다고 보는 것은 아니다. 책무성 기반 검증에서 투명성은 필요한 요소 중 하나이며(Kaminski, 2020), 알고리즘의 공정성과 관련된 중요한 부분이다. 그러나, 사례를 통해 문제제기를 하고자 하는 이유는 알고리즘의 투명성과 공정성은 단순히 알고리즘의 내용을 설명하거나 알고리즘을 바꾸는 것으로 끝나는 문제가 아니라 알고리즘을 끊임없이 감시해야 하고 그 결과를 추적할 필요성이 있다는 점을 이야기하기 위해서이다. 사례에 대한 배경에서 상세히 설명하겠으나, 본 사례는 아나니와 크로포드(Ananny & Crawford, 2018)가 지적한 투명성의 한계를 정확히 보여주면서 책무성 기반 알고리즘 검증이 필요하다는 것을 방증하는 사례가 될 것이다.

보다 구체적으로 본 연구를 통해 화두를 던지고자 하는 질문은 다음과 같다. 네이버에 공문을 보낸다고 해서 네이버가 본인들이 운영하는 알고리즘을 변경하거나 결과치를 아무런 설명 없이 바꾸는가? 알고리즘이 공정하다는 전제하에 공정하게 짜여진 알고리즘이라 하더라도 누군가가 공문을 보내면 조용히 수정 운영하면서 아무런 언급없이 알고리즘이 운영되는 불투명성을 보여주는가? 투명성과 관련된 질문이다. 알고리즘 책무성과 관련하여서는 알고리즘이 보여주는 결과에 대한 사회적 규범과 합의에 따른 평가가 아니라 자의적인 변경을 진행하는지에 대한 질문을 제기할 수 있다. 사회적인 논의가 아닌 알고리즘에 의해 나타나는 특정한 결과만을 플랫폼이 임시방편으로 바꾸면 "공평한 서비스", "깨끗한 서비스", "깨끗한 검색 결과"로 여길 수 있는가?

2) 연구 방법론

알고리즘 책무성을 어떻게 검증할 것인가, 특히 결과를 두고 검증하는 방법은 학술적으로 정의된 바가 없다. 본 연구에서는 인터넷 웹페이지의 시간상 전후 관계를 비교하여 그 변화의 원인이 무엇인지 추정하는 역설계 방식을 활용하였다(King, Pan, & Roberts, 2013). 이는 저널리즘 분야의 알고리즘 책무성 보도(Algorithmic Accountability Report)에서 활용한 방법이기도 하다(예, Diakopoulos, 2015). 저널리스트 관점에서 알고리즘은 블랙박스이기에, 역설계 방식으로 투입과 산출 간의 특정한 관계를 확인하는 것이다. 예컨대, 쇼핑몰에서 같은 제품의 가격을 이용자마다 다르게 책정하고 있음이 포착되면, 의심되는 요인들을 체계적으로 차별화하여 이용자 프로필 여러 개를 설정한 후 가격에 영향을 미치는 요인을 찾아낼 수 있다. 이렇게 검증이 필요한 알고리즘 현상을 발견하면, 알고리즘에 투입된 데이터와 산출된 결과 간의 관계를 확인할 수 있는 표본을 확보하고 검증하는 과정을 거쳐 뉴스 스토리로 만들어 보도하는 것이다(Diakopoulos, 2015). 이 방식은 또한 중국 정부의 소셜미디어 검열과 관련하여 사용된 바 있는데, 검열이 일어나기 전에 포스팅을 수집하고 검열이 일어난 이후의 포스팅을 수집하여 비교한 후, 어떠한 원리로 포스팅이 사라졌는지 검증하는 방식으로 적용되었다(King et al., 2013).

본 연구도 이를 차용하여 두 가지 방식의 표본을 수집하였다. 같은 검색어가 특정한 사건 전과 후에 각각 어떤 결과를 내는지 살펴보았고, 특정한 사건 발생 후에 서로 다른 검색어가 같은 검색 결과를 내는 경우도 탐색하여 수집하였다. 이 방법을 통해 특정한 사건 전과 후에 네이버 이미지 검색 알고리즘이 어떤 방식으로 변경되었는지 추론할 수 있는 근거 자료를 확보하였다.

4. 사례 분석: 네이버의 알고리즘 변경 사례

1) 배경: 시민단체의 문제제기

2022년 6월 시민단체 〈정치하는엄마들〉은 포털 사이트에서 일상적인 단어를 검색해도 "성적이고, 성편향적이며, 성차별적인 이미지"가 노출되는 것에 대해 문제를 제기하면서 문제 검색어를 제보받는 캠페인을 벌였고(정치하는엄마들, 2022, 6, 8), 일부 언론에서 캠페인 내용을 보도하였다(한예섭, 2022). 2015년 구글에서 일부 한글 검색어의 이미지 결과가 선정적이었던 문제와 유사한 상황이었다(김충령, 2015). 연구진은 이러한 보도자료 발표 이전부터 그와 같은 현상을 주시하면서 관련 검색어와 이미지 데이터를 수집하고 있었기에 〈정치하는엄마들〉의 활동과 포털 사이트의 반응에 주목할 수밖에 없었다. 네이버는 단체가 예시로 언급한 단어들의 이미지 검색 결과를 바로 변경했고, 관련하여 아무런 입장도 표명하지 않았다. 〈정치하는엄마들〉에 따르면 캠페인 내용을 보도한 뉴스 댓글에는 "지금 검색하니 그렇지 않은데 무슨 뉴스 기사인가라는 댓글이 달렸다"고 할 정도로 검색 결과 수정은 빠르게 이루어졌다. 몇 달 후 〈정치하는엄마들〉은 캠페인을 통해 수집된 문제의 단어들을 명기하여 검색어와 이미지 삭제를 요청하는 공문을 네이버를 포함한 포털 사이트에 보냈고(정치하는엄마들, 2022, 9, 5), 네이버는 이 중 일부 단어들의 이미지 검색 결과를 또 즉각 수정했지만 여전히 어떠한 반응이나 설명이 없었다.

2) 데이터

네이버 검색은 검색어를 조회하고, 이미지 결과 페이지에서 스크롤을 내리면 최대 550개 이미지 파일이 검색된다. 모든 검색은 로그아웃 상태에서 진행했고, 이미지는 세 가지 방법으로 수집했다. 첫 번째 방법은 검색 결과 화면의 스크린샷을 찍어 저장하는 것이었다. 이 방법은 전체 550개를 한 화면에 담을 수 없다는 단점이 있지만 가장 단순하고, 검색 결과를

즉각적으로 확인할 수 있는 장점이 있다. 본 연구진이 가장 중점적으로 수집한 방법은 두 번째 방법으로, 헤드리스(headless) 브라우저라 불리는 셀레니움(Selenium) 웹드라이버를 활용하여 검색 결과 사진들을 그대로 스크래핑하여 저장하는 것이었다. 이 방법을 사용하면, 네이버 이미지 검색 결과에 나온 최대 550개의 사진을 그대로 저장할 수 있을 뿐만 아니라, 해당 사진의 출처 URL과 사진이 검색되는데 활용되는 제목줄 문자열인 텍스트 태그(tag)도 수집이 가능하다. 마지막으로 네이버 개발자센터에서 제공하는 REST API라 불리는 네이버 API를 활용하였는데, 이 방법으로는 최대 1,100개까지 검색 결과 이미지를 수집할 수 있었다. 다만, 스크린샷의 경우 최근 구글이 학교별 G-Suite 데이터 서비스 용량을 줄이는 정책과 결부되어 데이터를 옮기는 과정에서 때로 유실되거나 미처 저장을 못하는 등의 이유로 모든 검색 결과 화면을 확보하지는 못했다. 그러나 두 번째 방법인 웹드라이버를 통한 이미지 파일과 출처 관련 정보 수집은 분석된 모든 검색어에 대해 완성된 상태에서 연구를 수행하였다. 즉, 아래 상술할 이미지 검색 결과의 모든 데이터는 스크린샷으로 저장한 '길거리' 데이터를 제외하고 웹드라이버로 스크래핑하여 구글 드라이브에 저장되어 있다.

〈정치하는엄마들〉이 맨 처음 보도자료를 낸 것은 2022년 6월 8일이었고, 이때 '길거리'와 '서양' 등의 단어가 예시로 포함되어 있었다. 이후 3개월여가 지난 9월 5일, 〈정치하는엄마들〉은 "원래 단어의 뜻과 상관없는 이미지 검색 결과가 나오는 검색어 예시"에 해당하는 단어 33개를 발표하고 이에 관한 공문을 구글과 네이버, 다음에 보냈다. 첫 보도자료가 발표됐던 시점에 연구진은 '길거리' 검색 결과 스크린샷만 확보하고 있었다. 이미지 검색 결과에 관심을 두고 있던 연구진은 '길거리' 스크린샷의 경우 2021년 9월 22일에 저장하였고, 과거 시점의 연도별 검색 결과도 모아놓고 있었다. 이때 수집해 저장된 이미지 검색 결과(스크린샷)는 2022년 5월에 실시한 검색 결과와도 동일하였으나, 보도자료 발표 이후 변경된

것을 확인했다. 공문에 포함된 다른 단어들의 이미지 검색 결과는 7월과 8월 중에 앞서 언급한 두 번째 방법으로 스크래핑하였다. 〈정치하는엄마들〉이 문제제기한 단어는 총 33개였고, 그중에서 연구진이 이미지 검색 결과를 수집한 단어는 26개였다(〈표 1〉).[2] 9월 공문 발송 이후 연구진은 해당 단어들에 대한 이미지 검색 결과를 다시 수집하였다.

전체 26개 단어 중 18개(69%)는 문제제기 전과 후의 변화가 없다고 판단되거나, 변화했다고 판단하기 어려웠다. 그러나 나머지 8개 단어는 공문발송 전과 후가 확연히 구분되는 결과를 발견했다. 이 글에서는 이들 8개의 사례를 모두 게시하고자 한다. 그 이유는 본 연구가 어떠한 변경이 이루어졌는지 그리고 변경의 방식이 어떠한지 논증하는데 있어서 데이터로 증거하는 것이 중요하다고 판단하기 때문이다. 본 논문에서 보도자료나 공문 발송 이전의 검색 결과는 웹드라이브에 저장되어 있는 스크린샷으로 제시하는 단어들이 많은데, 네이버가 검색 이미지 결과를 바꾸기 전에 수집한 내용을 저장한 것이다. 웹드라이브 스크린샷은 네이버 검색 결과 화면을 직접 찍은 스크린샷의 이미지들과 다르지 않다는 점을 언급하고자 한다. 마지막으로, 본 연구가 학술적 목적에서 수집한 이미지들을 제시하고 있기는 하나 검색 결과에서 인물을 특정할 수 있는 얼굴 등의 정보가 포함되는 것이 바람직하다고 볼 수 없어, 해당 경우는 본래 수집한 사진을 흐리게 보이는 블러(blur) 처리를 하였다.

표 1. 분석 대상 검색어: 네이버 이미지 검색

꼭지	꽐라	길거리	다리	도끼	돌핀팬츠
디스코팡팡	레이싱	레전드	만취	모델	민폐
사진집	아찔	여신	영계	움짤	일러스트
의상	자국	조공짤	직캠	코스프레	피지컬
호불호	혼혈				

2) 연구진은 1차 보도자료가 공지된 이후 〈정치하는엄마들〉에 접촉하여 각자의 입장을 공유하는 자리를 가졌다.

3) 분석: 이미지 검색 결과의 변화

(1) 복합 검색어로의 검색어 대체

① 길거리

본 연구진은 2021년부터 '길거리' 단어에 관심을 두고 있었고, 당시 저장한 '길거리' 검색 결과는 〈그림 1a〉와 같다. '길거리'는 〈정치하는엄마들〉이 캠페인 관련 첫 번째 보도자료를 냈을 때(2022년 6월 8일) 언급된 단어로, 이 보도가 이루어진 후 변화된 검색 결과는 〈그림 1b〉와 같다. 이 결과를 보면, '길거리'에는 남성도, 겨울도, 아이도, 노인도 없는데, 이러한 현상은 이지은(2020)이 이미 논의한 바 있다. 그런데, 흥미로운 점은 언론 보도 이후 바뀐 이미지 결과의 일관된 패턴이다. 모두 하나같이 양쪽에 건물이 늘어서 있고, 그 가운데 도로가 소실점을 향해 뻗어있는 구도의 이미지로 대체되었다는 점이다. 연구진은 〈그림 1b〉에 나온 획일적인 모습의 출처를 찾아 제목 태그를 관찰하여, 모든 태그에 '사진'이라는 단어가 공통으로 들어있는 것을 확인했다. 그 뒤, 네이버 검색창에 '길거리 사진'이라는 검색어를 입력했을 때 결과는 〈그림 1c〉와 같다. 〈그림 1b〉와 〈그림 1c〉를 비교하면 확연히 드러나듯 변경된 '길거리' 검색 결과는 '길거리 사진'이라는 검색 결과와 동일하다. 다시 말해, 기존의 '길거리' 이미지 검색 결과를 변경하기 위해 '길거리'라는 검색어에 '사진'이라는 단어가 추가되어 검색되도록 수정한 것으로, 복잡한 방식으로 특정한 알고리즘을 변경해서 혁신한 것이 아니라 복합 검색어를 활용하도록 알고리즘을 변용한 것이다. 그 결과 검색된 결과 사진의 위치도 같다.

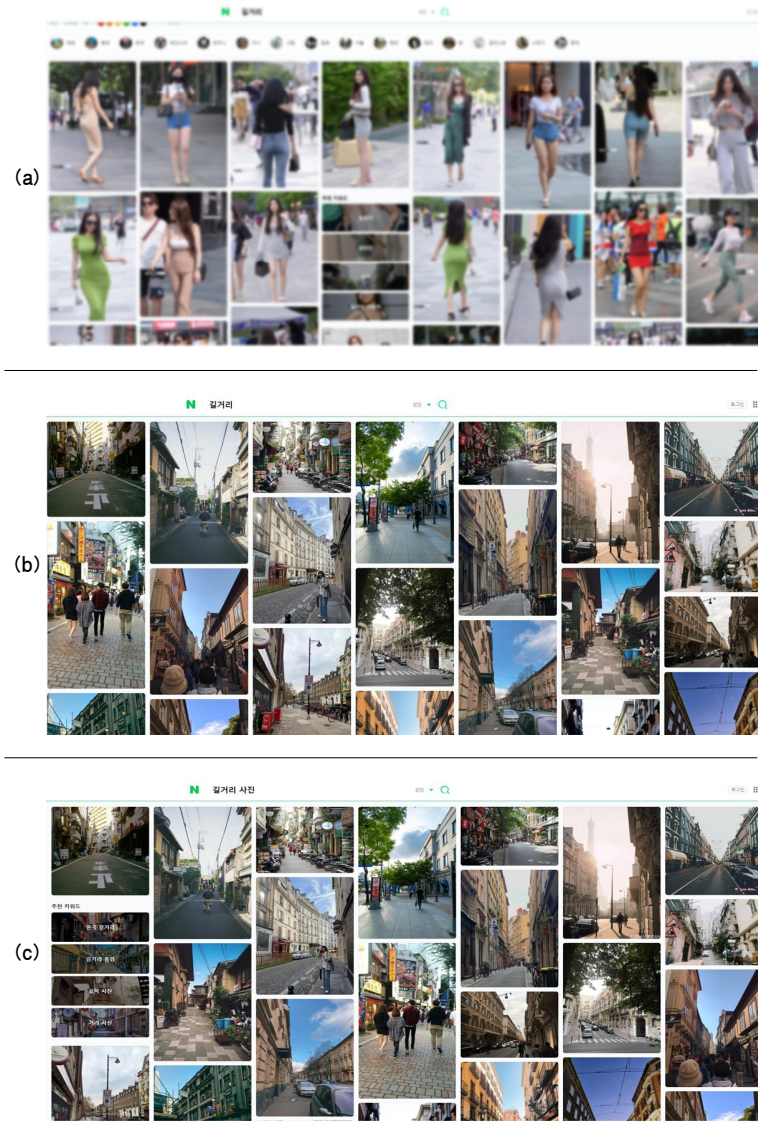
② 레이싱

'레이싱' 검색 결과도 위와 같은 방식으로 변경되었다. '레이싱'을 검색하면 원래 레이싱 모델이 주가 되어 나오던 이미지 검색 결과를 감추고, 다

른 결과를 보여주기 위해 네이버가 선택한 방식은 알고리즘을 개선하여 바꾸는 것이 아니라 '레이싱 경기'라는 다른 복합 검색어로 대체하는 것이었다. 〈그림 2a〉는 본 연구진이 공문 발송 전에 수집한 이미지들이고, 〈그림 2b〉는 그 이후 검색된 결과이다. 〈그림 2c〉는 '레이싱 경기'라는 '레이싱 AND 경기'의 복합 검색어로 이미지 검색을 한 결과인데, 〈그림 2b〉에 나온 이미지 검색 결과와 일치한다. 검색 결과에 대한 문제제기를 복합어 검색 결과로 대체함으로써, 원래의 문제적인 이미지를 문제가 없어 보이는 특정한 종류의 이미지로 바꾼 것이다. 알고리즘이 작동하는 방식을 개선하는 것과는 거리가 먼 이른바 원포인트 수정이라 부를 수 있는 방식이다.

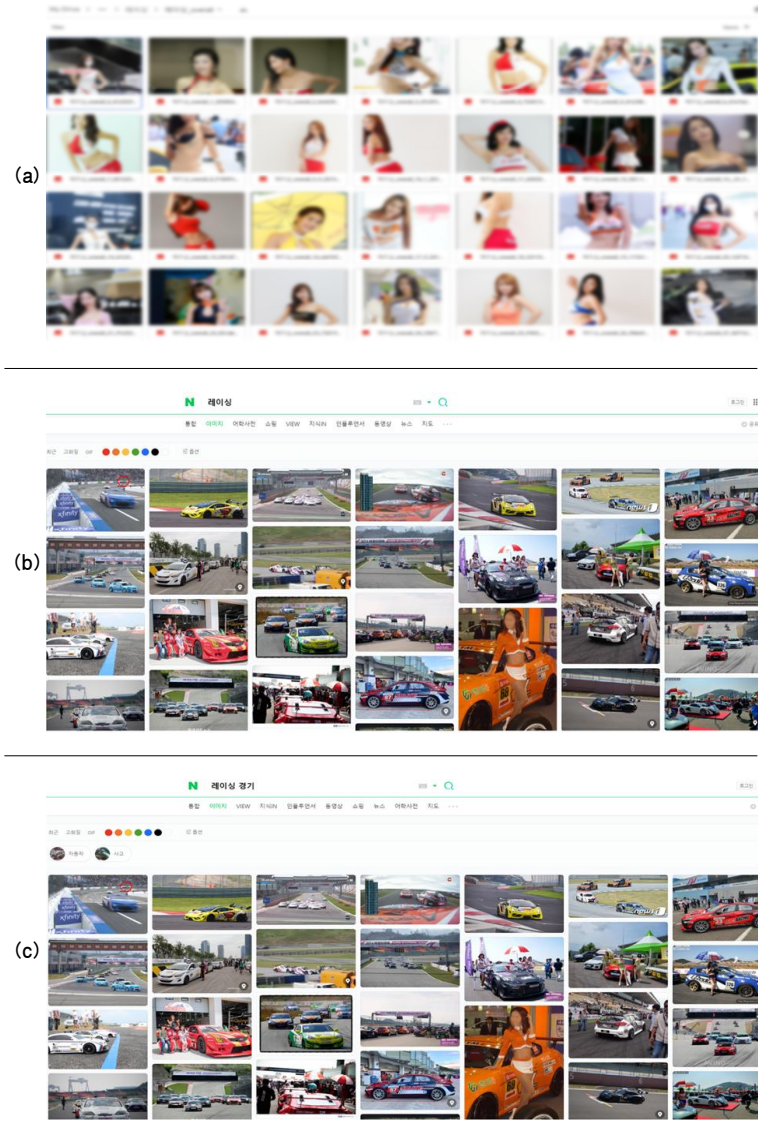
③ 모델

복합어로 대체하는 방식은 이에 그치지 않는다. '모델'이라는 검색어를 입력하면 이미지 결과는 본래 속옷만 입은 여성 사진이 대부분인 이미지 검색 결과가 나오고 있었다(〈그림 3a〉). 그런데 이러한 검색 결과를 사라지게 만드는 방식은 '패션 모델' 검색 결과가 나오도록 변경하는 것이었다. 앞의 두 사례와 마찬가지로 〈그림 3b〉와 〈그림 3c〉는 현재의 '모델' 단어에 대한 이미지 검색 결과가 사실은 '패션 모델'에 대한 이미지 검색 결과와 다르지 않음을 보여준다.



(a) 문제제기 이전(2021년 10월, 흐리게 처리함), (b) 문제제기 이후(2022년 9월), (c) '길거리 사진' 검색 결과(2022년 9월)

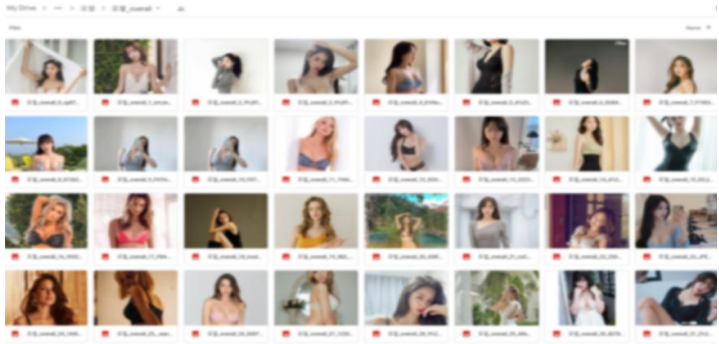
그림 1. '길거리' 네이버 이미지 검색 결과



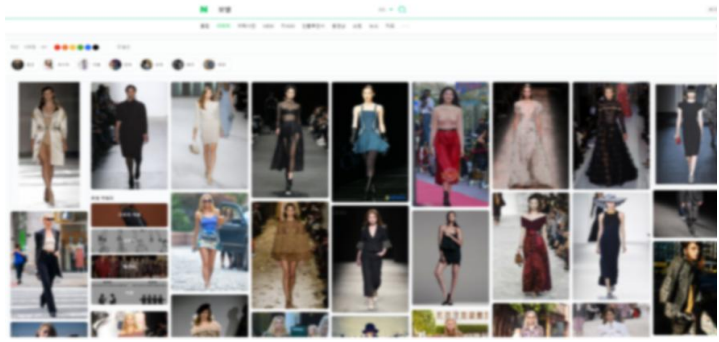
(a) 문제제기 이전(2022년 6월, 흐리게 처리함), (b) 문제제기 이후(2022년 11월), (c) '레이싱 경기' 검색 결과(2022년 11월)

그림 2. '레이싱' 네이버 이미지 검색 결과

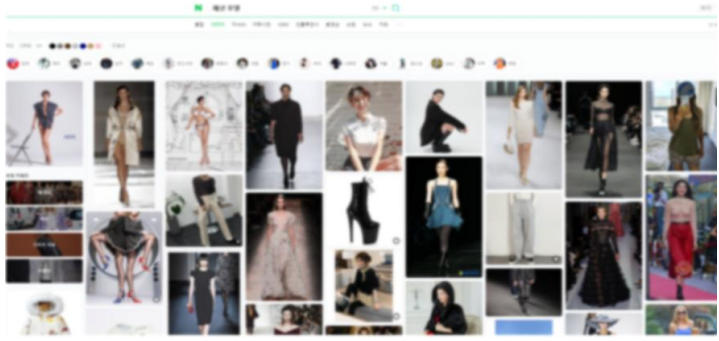
(a)



(b)



(c)



(a) 문제제기 이전(2022년 6월), (b) 문제제기 이후(2022년 11월), (c) '패션 모델' 검색 결과(2022년 11월) 모두 흐리게 처리함

그림 3. '모델' 네이버 이미지 검색 결과

(2) 유사어 검색 결과로의 대체

두 번째 변경 패턴으로 보이는 방법은 유사어 검색 결과로 대체하는 것이었다. 이는 검색 결과에 나오는 이미지들을 아예 다른 검색어 이미지 결과로 조정하여 바꾸는 것이다. 연구진은 이러한 경우를 복합어 검색보다 조정자의 의도나 편향이 더 많이 반영될 수 있는 변경 방식이라고 판단한다.

④ 사진집

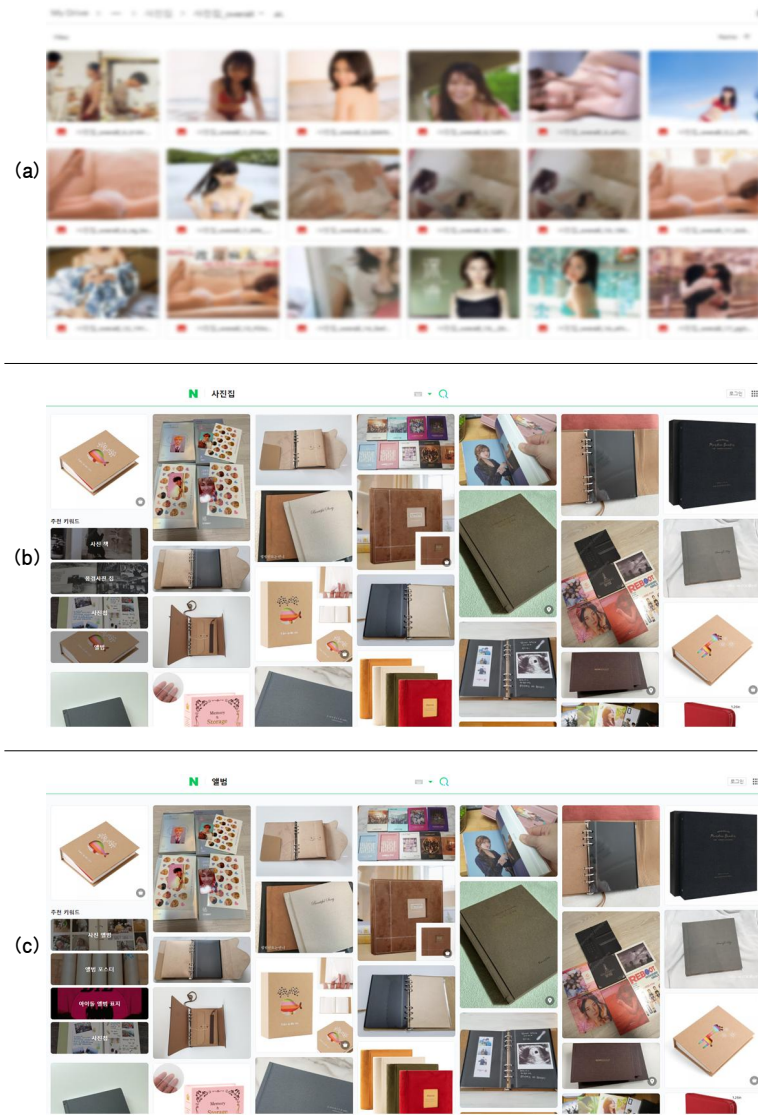
〈그림 4〉를 보면 문제제기 전과 후의 결과가 확연히 다른 것을 알 수 있다. 변경 후 '사진집'의 이미지 검색 결과는 물리적인 앨범 사진들이 검색되는데(〈그림 4b〉), 이러한 검색 이미지는 '앨범'이라는 단어로 대체한 결과를 보여주는 것에 불과하다 것을 알 수 있다(〈그림 4c〉). 다시 말해, '사진집'이라는 검색어와 '앨범' 검색어의 이미지 검색 결과가 동일하다. 이와 같은 변화를 추적하는 것은 앞서와 마찬가지로 어렵지 않았는데, 왜냐하면 바뀐 검색 결과 이미지에서 문자열 제목 태그를 쉽게 확인할 수 있었기 때문이다. 알고리즘의 개선이 아니라 단순 '교체'였던 것이다.

본 연구진처럼 관심을 가지고 추적하지 않은 상태의 일반 사용자들은 이러한 사실이 존재하는지는 전혀 모르고 변경된 이미지 결과를 접하게 된다. 특정한 검색 결과가 나오는 이미지를 교체하기 위해 다른 단어에 대한 검색 결과로 대체함으로써, 플랫폼은 알고리즘 운영의 투명성과 책무성을 벗어나게 되는 사례가 된 것이다.

⑤ 아찔

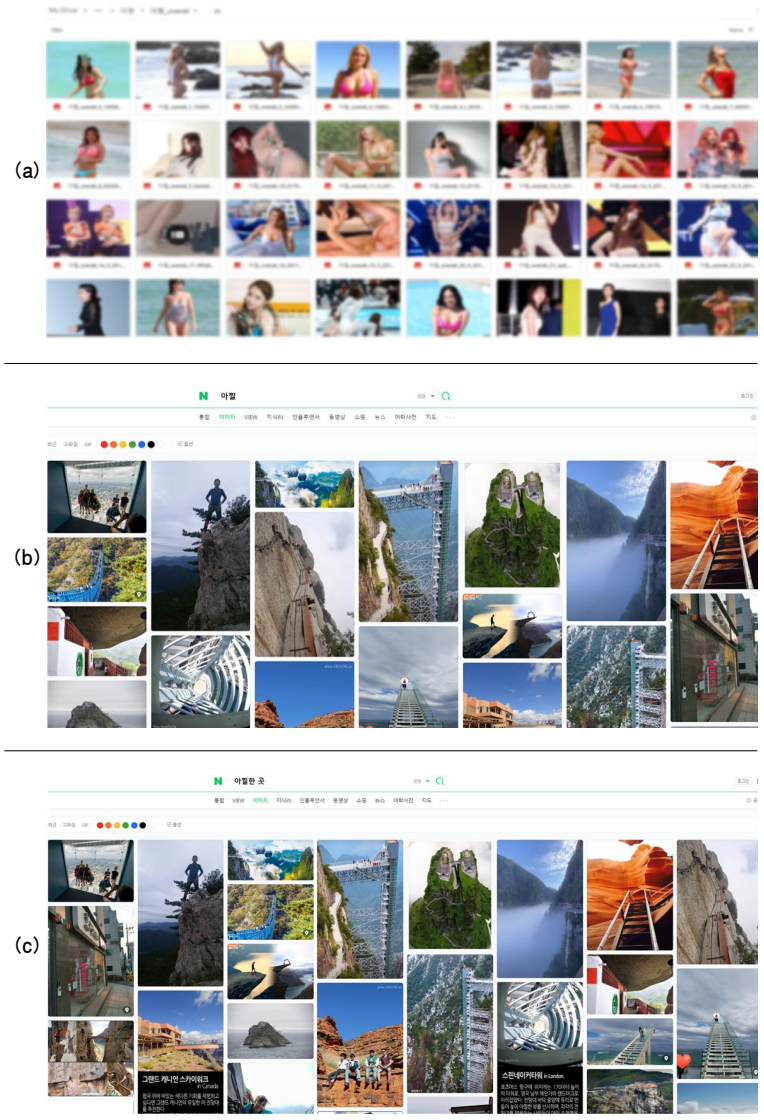
'아찔'이라는 단어도 유사한 단어로 대체되었다. 다만, 사진집과는 다소 다른 방식인 '아찔한 곳'이라는 검색어로 대체되었다. '아찔'이라는 단어가 '아찔한 곳' 단어로 대체되어서 복합어로 분류할 수도 있겠으나 특정한 독립적 단어가 결합한 것이 아니라 '부사'를 동사 활용형으로 바꾼 후 이에 대한 수식을 받는 의존명사를 등장시킨 형태다. 따라서 이는 복합어 검색

결과로 교체한 것이라기보다는 유사어 단어 대체로 분류하였다. 아래 〈그림 5〉에서 앞서 예시들과 마찬가지로 공문 발송 이전, 그 이후, 그리고 대체된 검색 결과 순으로 제시하였다. 〈그림 5a〉에서 보듯 원래는 비키니를 입은 여성 중심의 사진이었는데, 변경된 결과는 산과 같이 높은 곳에서 아래를 내려다 보는 장면의 사진들이 주를 이룬다(〈그림 5c〉). '아찔'의 검색 결과는 '아찔한 곳'의 검색 결과로 대체되었다.



(a) 문제제기 이전(2022년 6월, 흐리게 처리함), (b) 문제제기 이후(2022년 11월), (c) '앨범' 검색 결과 (2022년 11월)

그림 4. '사진집' 네이버 이미지 검색 결과



(a) 문제제기 이전(2022년 6월, 흐리게 처리함), (b) 문제제기 이후(2022년 11월), (c) '아찔한 곳' 검색 결과 (2022년 11월)

그림 5. '아찔' 네이버 이미지 검색 결과

(3) 자동완성 검색어 대체

세 번째 대체 방식은 자동완성 검색어를 활용하는 것이었다. 이 경우는 유사 검색어를 찾기보다는 이미 기존에 있는 문자열 태그를 그대로 활용하는 방식이라 보면 된다. 굳이 새로운 단어 조합을 만들거나 유사어를 찾기보다 데이터가 반영하는 부분을 그대로 활용하는 것이다. 물론, 이미지 검색 결과를 의도적으로 바뀌기 위해 데이터를 활용하는 것이지, 데이터를 그대로 반영하는 것이 아니라는 점에 주목할 필요가 있다. 아래에서는 '호불호', '레전드', '피지컬'의 세 단어에 대한 결과를 제시하고자 한다. 이 단어들은 중립적인 의미의 단어임에도 불구하고 이전의 검색 결과는 여성을 성적 대상화한 사진들이 주를 이루었다. 이 문제를 처리하기 위해 교체된 검색어는 자동완성기능이 생성하는 목록의 1순위에 나오는 검색어이다(〈그림 6〉).

(a) '호불호' (b) '레전드' (c) '피지컬'

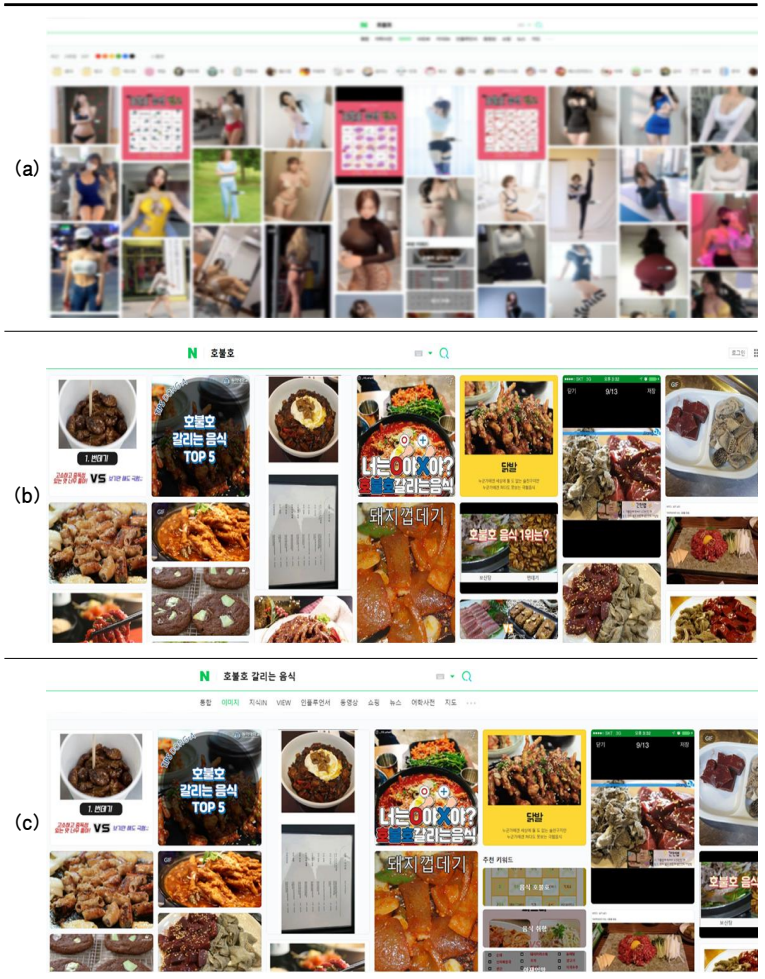
그림 6. 네이버의 자동완성 기능 예시(2022년 11월)

예를 들어, '호불호'를 입력하면 '호불호가 갈리는 음식'이라는 어구가 자동완성 검색어 목록의 가장 상단에 나오는데(〈그림 6a)〉. '호불호'의 이미지를 입력하면 자동완성기능으로 생성되는 검색어들 중에서 1순위에 해당하는 '호불호가 갈리는 음식'의 이미지로 대체되는 것이다.[3] 이 단어

3) 네이버 고객센터에 따르면, 자동완성 서비스는 이용자가 찾으려는 내용을 검색어로 최대한 잘 표현하게 도와주는 역할을 하고, 사용자의 검색 패턴 및 주요 정보 반영으로 인해 노출되는 검색어가 수시로 변할 수 있고, 시스템에 의해 자동으로 반영된다. https://help.naver.com/service/5627/contents/4955?lang=ko

들의 이전 검색 결과, 그리고 바뀐 결과, 대체된 검색어에 따른 이미지 검색 결과는 다음과 같다.

⑥ '호불호'

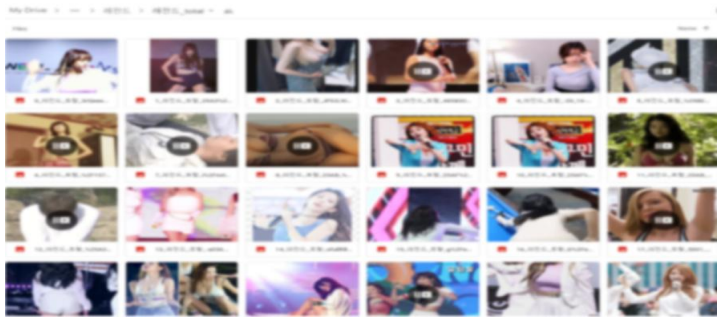


(a) 문제제기 이전(2022년 6월, 흐리게 처리함), (b) 문제제기 이후(2022년 11월), (c) '호불호 갈리는 음식' 검색 결과(2022년 11월)

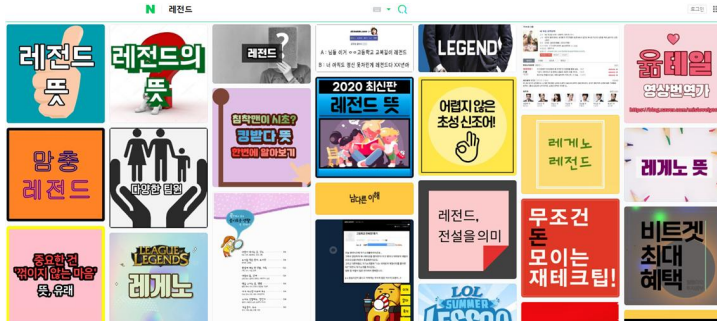
그림 7. '호불호' 네이버 이미지 검색 결과

⑦ 레전드

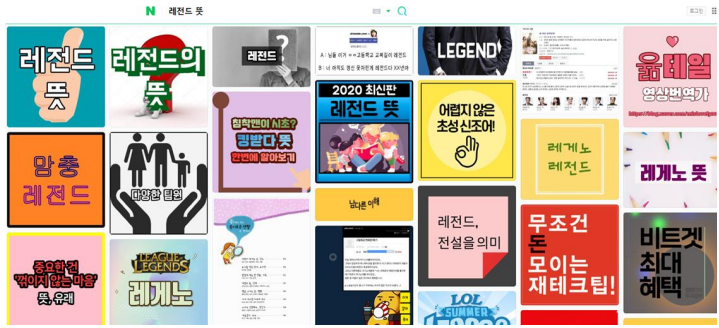
(a)



(b)



(c)

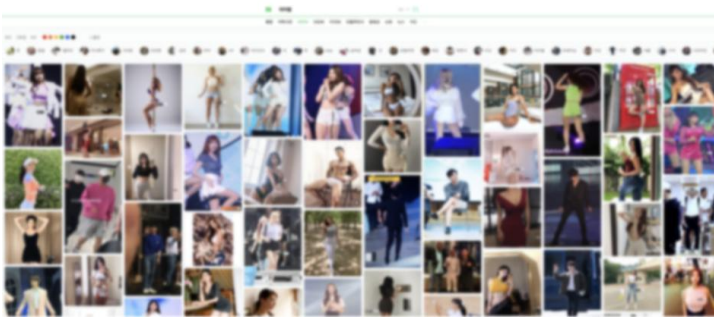


(a) 문제제기 이전(2022년 6월, 흐리게 처리함), (b) 문제제기 이후(2022년 11월), (c) '레전드 뜻' 검색 결과 (2022년 11월)

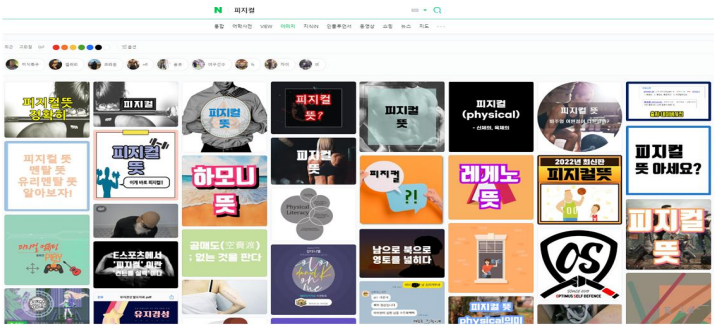
그림 8. '레전드' 네이버 이미지 검색 결과

⑧ 피지컬

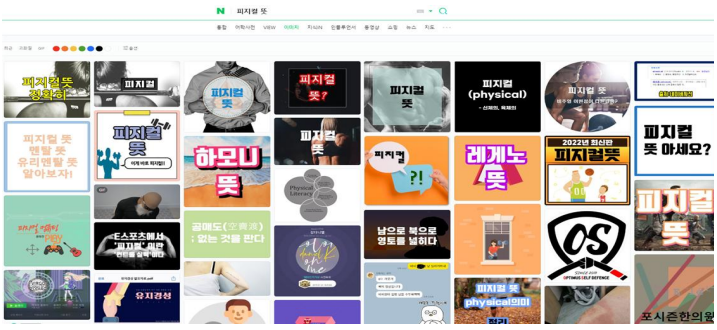
(a)



(b)



(c)



(a) 문제제기 이전(2022년 6월, 흐리게 처리함), (b) 문제제기 이후(2022년 11월), (c) '피지컬 뜻' 검색 결과 (2022년 11월)

그림 9. '피지컬' 네이버 이미지 검색 결과

이상과 같이 시민단체가 제기한 문제 검색어 목록에 있던 단어들 중 일부는 이미지 검색 결과가 변경되었다. 연구진이 수집하여 검증한 이미지 검색 결과의 변경 방식은 검색어에 따라 해당 단어가 포함된 복합 검색어로 대체, 유사 단어(어구)로 대체, 자동완성 서비스가 제공하는 검색어로 대체로 유형화할 수 있다. 연구진은 이미지 검색 결과를 수집하고 분석하는 동안 네이버가 검색 알고리즘을 개정하거나 개선했다기보다는 개별 검색어마다 보기 불편한 이미지들을 적당히 가리기만 한 것으로 파악했다. 문제가 드러나지 않도록 적당한 키워드를 골라 대체하는 방식으로 임시변통의 땜질을 한 것으로 본다.

플랫폼 측은 아마도 이와 같은 방식이 아니면 '전체 알고리즘을 수정하는 방식으로 개선이 이루어져야 한다'고 언급하거나 혹은 '알고리즘 개선 조치를 수행하는 과정 중에 우선 해당 검색어들의 결과에 대한 임시 조치를 취하는 것이 옳다고 판단했다'든가 '개선에 시간이 걸리기 때문' 등등의 언급을 할 것으로 예상된다. 물론, 알고리즘 대체에 걸리는 시간이나 알고리즘 자체에 대한 개선도 중요하겠지만, 본 연구에서는 그러한 기술적 내용이 아니라 운용상의 투명성과 책임성 등 보다 근본적인 문제를 제기하는 것이다. 본 연구진이 진행하고 있는 또 다른 연구에 따르면 현재의 이미지 검색 결과가 나타나는 원인은 의외로 쉽게 파악 가능한 측면도 있다고 보기에, 본 연구진은 '알고리즘의 개선'과 관련된 기술적 문제는 실제 투여하는 자원의 크기와 관련이 있다기 보다는 사실상 사회적 압력의 크기와 비례한다고 본다. 뉴스 알고리즘의 공개와 개선 시도가—정확히 일치하는 사례인지 그리고 해당 방식이 옳다고 볼 것인가는 논쟁의 여지가 있으나—주로 사회적인 요구와 압력에 의해 이루어졌다는 점은 참고할 필요가 있다.

5. 토론: 제기되는 질문들

본 연구에서는 알고리즘의 변경이 특정한 원리 원칙에 의해 이루어진 것이 아니라 자의적으로 아무런 설명없이 이루어지고 있는 사례를 발견했다. 시민단체가 제기한 네이버 이미지 검색 결과의 선정성 문제에 대해 네이버는 즉각 대응하였는데, 네이버는 제기된 검색어 중 일부만 수정하였고 이러한 선택과 조치에 대해 아무런 설명도 하지 않았다. 이 사례를 통해 본 연구는 알고리즘의 투명성 못지않게 플랫폼의 알고리즘 운영의 투명성과 알고리즘에 의해 나타나는 결과에 대한 책무가 중요하다는 점을 강조하고자 하였다.

본 연구에서 제시한 예시들을 살펴보면, 네이버는 이미지 검색 결과로 나타나는 결과를 일부 변경하였는데, 전체적인 검색 알고리즘을 변경한 것이 아니라 콘텐츠의 가시성(visibility)을 조절하는 방식으로 변경하였다. 입력된 검색어를 다른 검색어로 대체하거나 필터를 사용하여 특정 결과가 검색에 걸리지 않도록 하는 것으로 일부 검색어에 한해서만 다른 결과가 나오도록 한 알고리즘 운용 변경이다. 본 연구에서 보여준 콘텐츠 조정 방법은 모두 검색어를 다른 검색어로 대체하는 방식을 쓴 것으로 보인다. 기술적으로 단순하고 깔끔하게 처리함으로써 은밀하게 그리고 적극적으로 콘텐츠를 조정한 것이다. 최근 이미지 검색 문제의 선정성에 대해 취재한 언론 보도가 있었는데, 인터뷰 내용에 따르면 네이버는 선정적인 이미지 결과가 나온 것은 예측할 수 없었던 결과이고, "기술적으로 어려움이 있지만 선정성과 관련 기술·운영 보완과 함께 제기한 문제는 일정 부분 해소될 것"이라고 말했다(장슬기, 2022). 비록 알고리즘의 불가해성을 확인해 주고 기술적 처리를 약속하였으나, 이미 처한 조치가 어떠한 방식으로 진행되었는지에 대해서는 언급하지 않았다. 우리는 지금까지도 네이버가 어떤 근거와 기준을 통해 이미지 결과들의 수정 여부 및 수정 내용을 결정했는지 알 수 없다. 또한, 실제로 알고리즘 변경이

있는 것인지 해당 사항만 바꾼 것인지 또한 알 수 없다. 〈정치하는엄마들〉의 문제제기와 이후 일련의 과정들, 그리고 이와 관련된 본 연구진의 자료 수집이 있지 않았다면 전혀 알지 못한 채 지나갔을 일들이다.

이러한 발견은 다양한 질문을 제기한다. 가장 우선적으로 만약 〈정치하는엄마들〉이 캠페인을 벌이지 않았다면, 이 단체의 보도자료를 언론이 보도하지 않았다면, 알고리즘은 수정되었을까라는 의문이다. 그리고 그 과정에 대한 언론의 취재가 없었다면 어떻게 하고 있는지 설명했을까라는 질문도 이어진다. 결과 수정 과정과 관련해서도 상당한 의문이 존재한다. 어느 시민단체가 공문을 보내면 알고리즘은 변경될 수 있는 것인가? 그렇다면, 누군가 다른 개인이나 집단이 문제제기하면 알고리즘은 언제든지 바뀔 수 있는 것인가라는 문제제기도 동반한다. 이 문제를 확장하면 뉴스 서비스에 대한 문제제기도 가능하다. 이미지 검색 결과에서 공문에 의해 검색 결과가 바뀌는데 뉴스 서비스는 그렇지 않다는 보장이 어디에 있는가와 같은 질문이다. 전체적으로 알고리즘을 적용하지만 특정한 부분만 바꾸는 것이 가능하다는 것도 알게 된 셈인데, 이미지 검색 외에 다른 서비스는 그렇지 않다는 보장은 어떻게 가능한가라는 질문이 뒤따른다. 수정이 이루어진 결과에 대해서도 질문이 제기된다. 예를 들어, 요청을 수용하여 수정되면 모든 문제가 해결되는 것인가라는 문제다. 이제 성적 대상화된 여성들의 이미지가 검색되지 않게 되었으니 애초 그러한 결과가 발생하게 된 원인이나 이유에 대한 규명은 필요 없는 것일까라는 의문이다. 이러한 문제제기 마저도 네이버가 먼저 설명한 것이 아니라 제삼자인 연구자들이 연구하는 과정에 발견한 것인데, 연구자들이 발견하지 못했다면 누구도 알 수 없는 문제가 되었을 것이기 때문이다.

관련하여 보다 근본적인 문제제기도 가능하다. 우선, 검색 결과 이미지들을 조정하고 바꾸는 건 누가 정하는가의 문제다. 본 연구에서 밝힌 바에 따르면 전체 알고리즘의 내용을 변경한 것이 아니라 특정한 결과가 나오도록 몇 개의 단어에 대해서만 조정한 결과로 나타나고 있다. 이는

검색 결과를 인위적으로 조작하는 것과 본질적으로 다르지 않다. 또한 이러한 변경에 대해 아무런 설명을 하지 않음으로써 원래부터 그런 결과가 나오는 것처럼 위장한 효과도 있다. 이 문제에 대해 인지하지 못한 사람들은 변경 자체를 알아차릴 수 없을 것이다. 이 과정에서 어떤 콘텐츠가 노출되고 노출되지 않는지를 누가 결정하는가에 대한 것보다 더 중요한 문제는 의도한 결과가 나오도록 인위적인 조정을 아무런 사회적 대화와 논의 없이 진행하는 것이 과연 알고리즘의 책무성을 약속한 플랫폼의 정당한 행위인가라는 점이다.

당연하게도. 즉각적인 조치는 사람들이 해당 문제를 인식하고 그에 대해 사회적 논의가 이루어지는 길을 차단한다. 물론 (누군가에게는) 불편하고 문제가 있다고 여겨지는 콘텐츠는 언제나 존재할 것이다. 게다가 문제가 될 수 있는 콘텐츠는 불법이 아닌 경우도 많고, 집단마다 평가를 달리할 수도 있으며, 정치적이거나 전복적 가치를 지닐 수도 있다 (Gorwa, Binns, & Katzenbech., 2020). 또한 이렇게 특정한 검색 결과를 가린다고 하여도, 본래의 알고리즘 검색에 의해 나타나는 콘텐츠가 인터넷상에서 사라지는 것도 아니다. 본 연구에서 밝히고 있는 이미지 검색의 변경과 같은 일이 나타날 경우 특정한 현상이 나타나는 연유와 그러한 결과에 이르기까지의 과정—예를 들어, 데이터의 문제—전체를 논의할 수 있는 기회가 사라진다는 점도 생각할 필요가 있다. 그리고 실제로 이러한 기회가 사라진 것으로 보인다.

두 번째로 본 연구가 밝히는 이미지 검색 결과의 인위적 변경과 관련하여, 플랫폼 거버넌스에 대한 문제제기도 던질 수 있다. 네이버는 뉴스 배열 알고리즘을 공개하고 위원회로부터 정치적 비편향성에 대한 판단이 이루어진 후, 자체 블로그(Naver Search & Tech)에 뉴스 추천 알고리즘에 관해 여러 게시글에서 자세히 설명했다. 스스로 설명 요구에 응답한 것인데, 〈정치하는엄마들〉이 제기한 문제에 대해서는 별다른 응답이 없다. 응답은 하지 않은 채 일부 검색어의 알고리즘은 변경했는데,

변경 사유 중 하나로 추측되는 점은 이미지 검색 결과가 검색어의 의미와는 전혀 관계없는 여성들의 몸으로 도배되는 것은 네이버의 게시물 운영정책에도 명시된 바, 네이버의 신뢰도를 해치는 것으로 판단했을 수 있다는 점이다. 네이버의 뉴스 배열에 관한 문제제기는 상대적으로 오래됐고, 언론도 여론도 정치권에서도 주목한 이슈여서 투명성 요구에 대한 검증을 수용하고, 성차별적 이미지 검색 결과 문제는 아는 사람만 아는 내용이라 투명성 조차도 의미 없는 것으로 판단했는지에 대한 의문이 생길 수 있다.

세 번째로, 알고리즘을 수정해서 콘텐츠를 조정하면 불법이거나 명백하게 유해한 콘텐츠를 쉽게 차단할 수 있지만 이것으로 플랫폼의 책무성이나 윤리적, 정치적 책임이 덜어지는가에 대한 문제제기가 가능하다. 알고리즘을 슬쩍 변경하여 문제를 해결하는 것은 플랫폼의 책무성이나 윤리적, 정치적 책임을 덜어주지 않고 오히려 플랫폼의 자의적인 콘텐츠 조정은 그와 같은 시도에 대한 의심과 더불어 불신을 확산한다. 앞서 언급한 것처럼 이미지 검색이 이러하다면, 뉴스 알고리즘 운용도 필요에 따라 변경되는 것은 아닌가와 같은 합리적인 의심을 불러일으킬 소지가 다분하다. 즉, 그동안 어떤 콘텐츠를 어떻게 조정해왔고, 조정하고 있는지에 대한 의문이 들지 않을 수 없다. 다른 서비스에 대해서도 다양한 문제제기가 가능해진다. 예를 들어, 어떤 종류의 콘텐츠 순위가 알고리즘에 의해 자동으로 상향 조정되고 어떤 콘텐츠는 하향 조정 또는 배제되고 있을지, 뉴스 배치 알고리즘이 약속한 대로 개선되었다면, 이 개선은 구체적으로 무엇을 어떻게 수정한 것인지 등이다. 더군다나 네이버의 경우는 공정거래 위원회에 의해서 네이버 쇼핑에서 알고리즘을 자의적으로 변경해왔었다는 판단을 받은 바도 있다.

6. 결론: 진일보한 알고리즘 책무성 실천의 문제를 위하여

알고리즘 윤리를 의사들의 윤리와 같은 전문적인 직업 윤리와 비교하면서, 미텔슈타트(Mittelstadt, 2019)는 행위자들 간 공통의 목표와 실천 방식의 부재, 그리고 역사와 규범의 부재가 알고리즘 윤리의 문제점이라 지적한다. 알고리즘은 대개 공공 영역에 해당하는 일을 사적인 맥락에서 사용될 수 있도록 민간 영역에서 개발되는데 개발자, 사용자 그리고 알고리즘에 의해 영향을 받는 이해당사자들의 목표가 근본적으로 동일하지 않다는 문제가 있다. AI를 비롯한 알고리즘 개발이 공공 서비스가 아니며 개발 자체에 공적인 역할이나 직업윤리가 확립되지 않았기 때문에, 플랫폼 또는 개발자들이 자신의 이익이 아닌 사용자들의 이익을 최대한 보장한다고 신뢰할만한 근거가 없고, 당사자들의 개인적인 신념이나 평판이 훼손되는 것에 대한 두려움 등 직업윤리가 아닌 개인적인 양심에 의존해야 하는 문제가 있다는 것이다(Mittelstadt, 2019). 예를 들어, 이미지 조정을 담당한 네이버 개발자나 팀은 논란이 커질 수 있다고 판단하거나 개인적으로 불쾌하다고 여긴 이미지들을 가려내어 수정했을 수 있다. 그런데, 이렇게 플랫폼이 자체적으로 콘텐츠를 조정하는 것은 조정자의 편향이 반영되어 특정한 종류의 콘텐츠를 주변화시킬 수 있다(홍남희, 2020). 물론 플랫폼에 의존하거나 개발자에 의존하는 방식으로 잘 운영될 수도 있다. 그러나 이 역시 개발 당사자들조차도 알고리즘 책무성과 관련된 역사와 규범이 부재한 문제가 있다. 예를 들어, 의학 분야의 경우 오랜 기간의 역사를 통해, 문화적 차이와 전문화된 영역에 차이가 있어도, 전문적인 직업 문화(도덕적 의무)를 공유하고 있는데, AI 알고리즘 개발과 관련된 전문-직업적인 표준은 존재하지 않는다. 판단 규준이 없다는 것이다. 선한 의도와 최선을 다한다는 희망으로는 충분치 않다. 미텔슈타트는 법적인 규제와 사회적 관심이 부족한 상황에서, 자기 모니터링이 실패하는 상황도 단순히 윤리적 또는 신뢰할만한 AI 알고리즘이라는 잘못된

확신을 만들어내는 위험 요소를 가지고 있게 된다고 지적하고 있다.

따라서 실제적으로 알고리즘 책무성을 실천하기 위해서는 사회적으로 문제제기된 사항에 대해서는 사회적으로 이야기하고 공론화될 수 있는 경로를 만들고 책무성을 수행할 필요가 있다. 네이버는 알고리즘 개편과 관련하여 개편 사실조차 알리지 않고 검색 결과를 자사에 유리하게 나오도록 했다는 의심 사례도 존재한다. 그리고 이때 네이버는 50차례의 개편이 있는데 5번의 개편만 문제 삼았다는 입장을 보인적이 있다(정철운, 2020). 개편의 투명성을 기반으로 한 책무성이 아니라 횟수를 기준으로 설명한 것이다. 최근 오픈마켓 분야와 관련한 플랫폼 자율 규제 기구가 출범하고 검색과 추천 서비스의 투명성 제고를 위한 자율 규체 원칙을 적용하겠다는 발표가 있었다. 알고리즘이 투명하게 설명될 수 있어야 한다는 점에 대해서는 사회적, 제도적 합의가 이루어지고 있으며, 알고리즘을 만들어 실행하는 주체인 플랫폼 기업 또한 동의하고 있는 셈이다.

알고리즘의 책무성은 알고리즘이 공개되고 원리가 설명되는 것으로 끝나지 않는다. 알고리즘으로 인해 발생하는 상호작용과 거래, 사건과 사고 등 과정과 결과를 살펴볼 필요가 있다. 문제가 발견된 알고리즘이라고 해서 마치 아무 일 없었던 것처럼 재빨리 그 결과를 바꿔버리는 것 또한 책임과는 거리가 멀다. 플랫폼은 알고리즘이 내는 결과와 그 결과의 변화를 설명할 책무가 있다. 이를 위해 어떤 조건에서 어떤 방식으로 설명 책무를 이행해야 하는지에 대한 논의를 시작해야 한다. 그리고 문제가 있는 알고리즘의 결과를 어떻게 판단하고 검증할 것인지에 대한 논의도 필요하다. 특히, 기술적인 설명 투명성으로 알고리즘에 관한 사회적 책무가 완수된다고 보기 어렵다는 점도 논의될 필요가 있다.

참고문헌

과학기술정보통신부 (2021, 5, 13). 〔보도자료〕 사람이 중심이 되는 인공지능을 위한 신뢰할 수 있는 인공지능 실현 전략(안). URL: https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=112&pageIndex=&bbsSeqNo=94&nttSeqNo=3180239&searchOpt=ALL&searchTxt=

김국배 (2022, 11, 3). '알고리즘 개편 고지했나' 네이버-공정위 끝까지 날선 공방…내년 1월 결판. 〈이데일리〉. URL: https://www.edaily.co.kr/news/read?newsId=03434166632522112&mediaCodeNo=257&OutLnkChk=Y

김충령 (2015, 7, 4). 한글 'ㄱ'과 알파벳 'A'… 구글 검색 결과는 딴판. 〈조선일보〉. URL: https://www.chosun.com/site/data/html_dir/2015/07/03/2015070304234.html

오요한·홍성욱 (2018). 인공지능 알고리즘은 사람을 차별하는가? 〈과학기술연구〉, 18권 3호, 153-215.

유진상 (2022, 5, 24). 방통위 '포털 뉴스 신뢰성·투명성 제고를 위한 협의체' 출범. 〈IT조선〉. URL: https://it.chosun.com/site/data/html_dir/2022/05/24/2022052401575.html

이중원 (2019). 인공지능에게 책임을 부과할 수 있는가?: 책무성 중심의 인공지능 윤리 모색. 〈과학철학〉 22권 2호. 79-104.

이지은 (2020). "한국여성의 인권에 대해 알고 싶으면, 구글에서 '길거리'를 검색해 보라": 알고리즘을 통해 '대중들' 사이의 적대를 가시화하기. 〈미디어, 젠더 & 문화〉, 35권 1호, 5-44.

장슬기 (2022, 11, 22). 포털 검색 성적 대상화 이미지 결과 개선되긴 했지만… 〈미디어오늘〉. URL: http://www.mediatoday.co.kr/news/articleView.html?idxno=307072

정소영 (2022). 유럽연합 인공지능법안의 거버넌스 분석: 유럽인공지능위원회와 회원국 감독기관의 역할과 기능을 중심으로. 〈연세법학〉, 39권, 33-65.

정철운 (2020, 10, 6). 네이버의 '검색알고리즘 조작'이 사실로 드러났다. 〈미디어오늘〉. URL: http://www.mediatoday.co.kr/news/articleView.html?idxno=209643

정치하는엄마들 (2022, 6, 8). 〔보도자료〕 정치하는엄마들 미디어감시팀 〈포털사이트 검색 이미지를 바꾸자!〉 캠페인 진행. URL: https://www.politicalmamas.kr/post/2342

정치하는엄마들 (2022, 9, 5). 〔보도자료〕 정치하는엄마들 미디어감시팀 구글·네이버·다음 등 포털 사이트에 '문제 검색어 및 이미지 삭제 요청'. URL: https://www.politicalmamas.kr/post/2484

채새롬 (2021, 7, 21). "뉴스 추천 알고리즘 공정한가요"… 네이버 답변은. 〈연합뉴스〉. URL: https://www.yna.co.kr/view/AKR20210721123700017

최창원 (2022, 9, 6). 택시 기사 반발에 '기밀' 내놓은 카카오, '배차 알고리즘' 문제 없었다. 〈BLOTER〉. URL: https://www.bloter.net/newsView/blt202209060004

한예섭 (2022, 6, 9). '길거리' 검색하면 '길거리 OO녀', 성차별적 이미지 쏟아진다. 〈프레시안〉. URL: http://www.pressian.com/pages/articles/2022060915584002743

홍남희 (2020). AI와 콘텐츠 규제: 자동화된 차단 기술의 문제들. 금희조·강혜원 (편), 〈AI와 더불어 살기〉 (295-325쪽). 커뮤니케이션북스.

황용석·정재선·황현정·김형준 (2021). 알고리즘 추천 시스템의 공정성 확보를 위한 시론적 연구. 〈방송통신연구〉, 116호, 169-206.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973-989.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.

ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3-44.

Bernard, Z. (2017, December 20). The first bill to examine 'algorithmic bias' in government agencies has just passed in New York City. *INSIDER*. Retrieved from https://www.businessinsider.com/algorithmic-bias-accountability-bill-passes-in-new-york-city-2017-12

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European Law Journal, 13*(4), 447-468.

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research, 71*, 1137-1181.

Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Administrative Law Review, 71*(1), 1-56.

Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv*. doi:10.48550/arXiv.1408.6491

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism, 3*(3), 398-415.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe*. Retrieved form https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Dougherty, C. (2015, July 1). Google photos mistakenly labels black people 'gorillas'. *The New York Times*. Retrieved from https://nyti.ms/2opE8CD

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580.

Edwards, L., & Veale, M. (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*, 18-84.

European Union. (2021). *What is the EU AI Act?* https://artificialintelligenceact.eu/

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics, 26*(6), 3333-3361.

Forssbaeck, J., & Oxelheim, L. (2014). The multifaceted concept of transparency. In J. Forssbaeck & L. Oxelheim (Eds.), *The Oxford handbook of economic and institutional transparency* (pp. 3-30). Oxford University Press.

Goldman, E. (2006). Search engine bias and the demise of search engine utopianism. *Yale Journal of Law and Technology, 8*, 188-200.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic

content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 1-15.

Heikkilä, M. (2022, October 4). The White House just unveiled a new AI Bill of Rights. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/2022/10/04/1060600/white-house-ai-bill-of-rights/

Internet Trend. (2022, December 15). *Search Engine*. Retrieved from http://www.internettrend.co.kr/trendForward.tsp

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389-399.

Johnson, D. G. (2021). Algorithmic accountability. *Social Philosophy and Policy, 38*(2), 111-127.

Kaminski, M. E. (2020). Understanding transparency in algorithmic accountability. In W. Barfield (Ed.), *Cambridge handbook of the law of algorithms* (pp. 121-138). Cambridge University Press.

Khalid, A. (2022, February 3). Democratic lawmakers take another stab at AI bias legislation. *Engadget*. Retrieved from https://www.engadget.com/wyden-algorithmic-accountability-act-2022-205854772.html

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review, 107*(2), 326-343.

Kroll, J., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633-705.

Liu, H. W., Lin, C. F., & Chen, Y. J. (2019). Beyond State v

Loomis: Artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology, 27*(2), 122-141.

Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: Explaining and justifying inscrutable algorithms. *Ethics and Information Technology, 23*(3), 253-263.

Meijer, A. (2014). Transparency. In M. Bovens, R. Goodin, & T. Schillemans (Eds.), *The Oxford handbook of public accountability* (pp. 507-524). Oxford University Press.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501-507.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1-21.

Naver. (2021. 11. 29). *NAVER-SAPI AI Report*. Retrieved from https://www.navercorp.com/value/research/view/15

Pasquale, F. (2019). *The second wave of algorithmic accountability*. Retrieved from https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/

Popper, K. (2002). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Schot, J., & Rip, A. (1997). The past and future of constructive technology assessment. *Technological Forecasting and Social Change, 54*(2-3), 251-268.

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM, 56*(5), 44-54.

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018, December). Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp.

303-310). ACM.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: Key problems and solutions. *AI & Society, 37*, 215-230.

Vedder, A., & Naudts, L. (2017). Accountability for the use of algorithms in a big data environment. *International Review of Law, Computers & Technology, 31*(2), 206-224.

Wieringa, M. (2020, January). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 1-18). ACM.

Wyden, R. (2022). *Algorithmic Accountability Act of 2022*. https://www.wyden.senate.gov/imo/media/doc/2022-02-03%20Algorithmic%20Accountability%20Act%20of%202022%20One-pager.pdf

투 고 일 자: 2023년 04월 05일
심 사 일 자: 2023년 05월 07일
게재확정일자: 2023년 05월 24일

Abstract

Does Naver Alter Search Results upon Request?

Transparency and Accountability of Algorithms

Ho Young Yoon
Assistant Professor, Division of Communication & Media, Ewha Womans University
Borae Jin
Assitant Professor, Department of Media and Communication, Joongbu University

The growing concerns about bias or the fairness of algorithms have sparked intense debates and controversies surrounding the transparency and accountability of algorithmic systems. In this study, we present compelling evidence of algorithm alteration by the platform company, NAVER. In mid-2022, a non-profit organization brought attention to the appearance of sexually objectified images of women appearing in image search results on portal sites for unrelated search queries. The organization publicly disclosed 33 problematic words. Due to our pre-existing interest in this matter, we closely monitored the responses from the companies involved. Promptly, Naver adjusted its image search results. We conducted a comparative analysis, presenting the image search results before and after the modifications for the words mentioned in the organization's announcement. Our findings shed light on several critical concerns regarding Naver's algorithm operation. Primarily, the algorithm alterations seemed arbitrary, with problematic search words being replaced by other relevant words or autocomplete suggestions. Also, Naver selectively addressed only a

portion of the 33 words, and further, Naver did not offer any explanations or responses to the organization or its users. These observations raise questions about the transparency and accountability of algorithms employed by the platform. We argue that platform companies should not only clarify the functioning of their algorithms but also disclose instances and methods of algorithm moderation. Furthermore, we emphasize the urgent need for an ethical framework that holds platform companies responsible and accountable throughout the entire process of algorithm development, implementation, and the resulting consequences.

KEYWORDS platforms, Naver, algorithm, transparency, accountability