과학철학 28권 1호 (2025) pp. 33-64

인공지능 시대 과학자의 도덕적 책임: 합당한 예견의 기준에 관하여*

한 혜 정†

이십여 년 전, 헤더 더글러스는 「과학자의 도덕적 책임」이라는 글에서 과학자는 자신의 연구가 초래할 수 있는 의도치 않은 결과를 예견할 도덕적 책임이 있으며, 어디까지 예견하는 것이 합당한 지는 과학 공동체의 기준을 따라야 한다고 주장하였다. 이 논문의 목적은 더글러스의 주장에 대한 비판적인 검토를 통해 인공지능 시대의 과학자에게 요구되는 도덕적 책임을 논하는 것이다. 이 논문에서는 오늘날 인공지능 기반 과학 연구의 특징을 살펴보며, 더글러스의 주장이 인공지능 기반 연구에서 발생하는 "다학제의 문제", "느린 기준 설정의 문제", "편향된 공동체의 문제"를 포착하지 못함을 지적한다. 또한 이에 대한 해결책으로, 첫째, 연구 프로젝트의 구체적 맥락을 중심으로 합당한 예견의 기준을 설정할 것, 둘째, 다양한 관점을 적극적으로 고려하는 다원주의적 공동체를 조성하기 위해 일부 과학자에게 동료를 동원할 책임을 부여할 것을 제안한다.

주요어 : 합당한 예견, 의도하지 않은 결과, 과학 공동체, 공동체 기준, 맥락 의존성, 다원주의, 헤더 더글러스

* 이 논문을 작성하도록 격려해 주신 그랜트 피셔(Grant Fisher) 교수님께 감사드린다. 또한 도움이 되는 제안을 주신 익명의 심사위원들께도 감사를 전한다.

† 제주대학교 윤리교육과 강사 (hhj29@kaist.ac.kr)

1. 들어가며

과학자가 도덕적으로 연구를 수행해야 한다는 주장에 반대하는 사람은 많지 않지만, 구체적으로 어떤 도덕적 책임을 져야 하는지는 의견이 분분하다. 특히 위조나 표절 같은 연구 부정행위를 저지르지 말아야 하며, 동물 실험을 수행할 때 동물의 고통을 최소화해야 한다는 연구 방법 차원에서의 주장은 폭넓은 동의를 받는 반면[1] 연구 결과가 불러올 파장에 대해서는 다양한 이견이 존재한다. 예를 들어, 사람을 살리기 위해 진행한 연구가 과학자의 의도와 달리 사람을 해치기 위해 이용되었다면 과학자는 이에 대한 도덕적 책임을 져야 하는가? 의도치 않았던 결과에 대해 도덕적 책임을 부과한다면 과학자의 자율성을 침해할 우려는 없는가?

21세기 과학철학의 거장 중 한 명인 헤더 더글러스는 이십여 년 전 「과학자의 도덕적 책임」(Douglas 2003)이라는 제목의 논문을 통해 비슷한 질문을 던졌다. 그는 과학자가 의도하지 않은 결과를 고려할 도덕적 책임이 있으며, 이러한 책임이 자율성과 적대적인 관계가 아니라고 주장했다. 이때, 그의 주장의 핵심은 의도하지 않은 잠재적 결과를 모두 완벽하게 예견해야 한다는 것이 아니라, 과학 공동체의 기준에 부합하는 "합당한 예견(reasonable foresight)"이 필요하다는 것이다. 이처럼 과학 공동체가 제공하는 합당한 예견의 기준을 이 논문에서는 더글러스의 "공동체 기준"이라고 부르겠다. 더글러스가 공동체 기준을 내세운 바탕에는 과학 공동체가 어디까지 예견하는 것이 합당한지에 대한 **단일한** 기준을 **신속하게** 제공할 것이라는 가정이 깔려있다.

이 논문의 목적은 더글러스의 주장을 비판적으로 검토하여 오늘날 인공지능 기반 과학 연구에서 과학자의 도덕적 책임에 대해 논하는 것이다.[2] 주지하다시피, 최근 인공지능 기술의 비약적인 발전은 일상생활

1) 연구 부정행위에 대한 논의는 최훈·신중섭 (2007), 이상욱 (2011) 등을, 동물 실험에 대한 논의는 최훈 (2009), 목광수 (2010) 등을 참조.

2) 의약품 규제 의사 결정 시 인공지능 사용에 대한 미국 식품의약국(Food and Drug Administration; FDA)의 지침 초안을 따라(FDA 2025), 이 논문에서는 "인공지능"을

뿐만 아니라 과학 연구에 큰 변화를 불러왔다. 특히 2024년 노벨물리학상과 노벨화학상 수상자가 각각 인공신경망 연구와 단백질 구조를 예측하고 설계하는 인공지능 모델을 개발한 공로로 선정되었다는 사실은 인공지능이 오늘날 과학 연구에 지대한 영향을 미치고 있음을 보여준다. 현재 인공지능은 연구 초기 단계에서의 가설 탐색 및 발견 과정에서 널리 활용될 뿐만 아니라(Clark and Khosrowi 2022; Duede 2023), 의약품 안전성 평가와 같은 규제 결정 과정에서도 중요한 도구로 부각되고 있다(EMA 2024; FDA 2025). 하지만 동시에, 부정확한 정보를 그럴듯하게 제공하는 환각(hallucination) 현상 등의 문제가 대두되며 인공지능에 대한 우려도 함께 커지고 있다(Hicks et al. 2024). 일부 전문가들은 인공지능이 인류에게 심각한 위협이 될 수 있다고 주장하며 거대 인공지능 개발의 잠정적 중단을 촉구하는 공개서한을 발표하기도 했다(Futre of Life Institute 2023).[3)]

이러한 상황에서, 인공지능 연구자 공동체는 인공지능 기반 과학 연구가 불러올 수 있는 의도치 않은 결과에 대하여 더글러스가 말한 “합당한 예견”의 기준을 제시할 수 있는가? 혹은 다양한 이해관계자가 관여하는 인공지능 기반 과학 연구에서 합당한 예견의 기준은 무엇을 중심으로 설정해야 하는가? 이 논문에서 나는 인공지능 기반 연구의 특징을 살펴보며 더글러스의 “공동체 기준”의 문제점 세 가지를 논할 것이다. 구체적으로는 첫째, 인공지능 기반 과학 연구에서는 다학제 간 협력, 특히 인공지능 연구자와 실험가의 협력이 중요한데, 이들 각자의 공동체는 합당한 예견에 대해 서로 다른 기준을 가질 수 있다. 이는 더글러스가 전제한 단일한 기준이 성립하지 않을 수 있음을 시사한다. 둘째, 더글러스는 과학 공동체가 합당한 예견의 기준을 신속하게 제공할 것이라 가정하지만, 인공지능 기반 과학 연구에서는 인공지능 기술의 발전 및 응용 속도가 너무 빠른 나머지 공동체가 충분한 논의를 거쳐

인간이 정한 목표를 위해 실제 또는 가상 환경에 영향을 미치는 예측, 추천, 결정을 내릴 수 있는 기계 기반 시스템으로 정의한다.

3) 삶의 미래 연구소(Future of Life Institute)에서 발표한 공개서한. 스튜어트 러셀(Stuart Russel)과 존 홉필드(John J. Hopfield)를 비롯한 인공지능 업계의 권위자 여럿이 서명했다.

기준을 정하는 속도가 이를 따라가지 못할 가능성이 있다. 셋째, 더글러스는 공동체 자체를 도덕적 평가의 기준으로 상정하지만, 공동체 자체가 편향될 수 있음을 간과하고 있다.

이러한 문제점을 해결하기 위하여, 나는 과학과 가치에 대한 과학철학 및 사회인식론 분야의 문헌에서 논의되는 맥락 의존성(context-dependence)과 인식적 다원주의(pluralism) 접근을 적용하여 다음과 같이 제안하고자 한다. 첫째, 의도하지 않은 결과에 대한 합당한 예견은 단일하고 추상적인 공동체 기준에 전적으로 의존하기보다는 연구 프로젝트의 구체적인 맥락을 중심으로 설정되어야 한다. 둘째, 과학 공동체는 단일한 기준을 제시하는 대신 다양한 관점을 적극적으로 고려해야 하며, 이를 위해 일부 과학자에게는 동료를 동원(mobilize)할 책임을 부여해야 한다. 명확하게 말하자면, 이 논문에서 일컫는 **과학자**란 화학자, 생물학자 등뿐만 아니라 인공지능 연구자 등을 포함하며, 인공물이 아닌 **인간** 과학자만을 의미한다.[4] 또한 이 논문의 목적은 과학자가 의도하지 않은 결과를 고려해야 한다는 더글러스의 주장 전체를 반박하려는 것이 아니라, 합당한 예견에 대한 수정된 기준을 제시함으로써 인공지능 시대의 과학에 더욱 부합하는 제안을 하려는 것이다.

이 논문은 다음과 같이 구성된다. 먼저 2절에서는 과학자의 도덕적 책임에 대한 더글러스의 주장이 등장한 배경과 그 핵심 내용을 개괄할 것이다. 이어 3절에서는 오늘날 인공지능 기반 과학 연구의 특징을 소개하고, 합당한 예견에 대한 더글러스의 공동체 기준이 갖는 세 가지 문제점을 지적할 것이다. 이 문제점들은 다학제의 문제, 느린 기준 설정의 문제, 편향된 공동체의 문제로 명명될 것이다. 마지막으로 4절에서는 연구 프로젝트가 수행되는 구체적인 맥락에 대한 강조와 다양한 관점을 적극적으로 고려하는 다원주의적 공동체에 대한 강조를 바탕으로 더글러스의 공동체 기준에 대한 대안을 제시할 것이다.

4) 인공물이 도덕적 책임의 주체일 수 있는지에 대한 질문은 이 논문의 범위를 넘어선다. 관련 논의는 고인석 (2012), 이상형 (2016) 등을 참조.

2. 의도하지 않은 결과에 대해 과학자는 어떤 책임을 지는가? 더글러스의 주장

과학자의 도덕적 책임에 대한 더글러스의 주장을 깊이 있게 이해하기 위해서는 먼저 해당 주장이 등장한 배경을 살펴볼 필요가 있다. 제2차 세계대전 이후 미국을 위시한 서구 사회에서는 과학의 자유에 대한 지지가 광범위하게 형성되었다. 전후 미국 과학정책의 기반이 된 『과학, 그 끝없는 프런티어』(Bush 1945) 보고서는 기초과학이 응용연구가 되고 이것이 사회적 혜택으로 이어진다는, 이른바 선형 모델을 제시하며, 사회적 필요에 대한 제약 없이 정부가 기초과학 연구를 전폭적으로 지원해야 한다고 주장했다. 또한 비슷한 시기에 발표된 「과학자와 사회적 책임」(Bridgman 1947) 성명은 과학자는 연구의 사용에 대해 도덕적 책임이 없다고 강조했다.[5] 여기서 과학자에게 책임을 지우는 것은 곧 과학의 자유를 침해하는 것과 동일시되었다.

이러한 전후의 인식에 도전이 없었던 것은 아니다. 20세기 중후반 동안 살충제 개발 등으로 인한 환경 문제나 유전자 재조합 기술이 초래할 수 있는 위험에 대한 우려가 불거지며 과학 연구의 윤리적 문제에 대한 고찰을 촉구하기도 했다. 그럼에도 이는 몇몇 예외적인 사례로 취급되었을 뿐, 과학자가 자신의 연구가 초래할 수 있는 사회적 영향에 대해 책임지지 않아도 된다는 견해는 20세기 후반까지 큰 영향력을 유지했다. 예를 들어, 미국 국립과학원은 1992년에 『책임 있는 과학』(NAS 1992)이라는 제목의 보고서를 발간하였지만, 위조나 표절 같은 연구 부정행위를 강조한 데에서 그쳤다.[6]

더글러스가 2003년 발표한 과학자의 도덕적 책임에 대한 논문(Douglas 2003)은 이러한 미국 사회의 주류 견해에 대한 반박이었다.

5) 히로시마 원자폭탄을 특히 염두에 두고 작성된 글이다.

6) 과학자의 책임과 자유에 대한 미국 사회의 인식 변화에 대한 자세한 논의는 Douglas (2021)를 참조. 해당 문헌에서 더글러스는 20세기 말까지 과학자의 도덕적 책임에 대한 기대가 상당히 제한적이었으나, 2000년대 이후 분위기가 변화하여 과학자가 더 자유로울수록 더 많은 책임을 져야 한다는 견해가 널리 확산했다고 설명한다.

더글러스는 이 논문에서 과학자는 본인의 연구가 미칠 사회적 영향에 대해 고려할 도덕적 책임이 있음을 주장했다. 해당 논문은 이후 21세기 과학철학에서 과학과 가치 논의의 핵심 저서로 꼽히는 『과학, 정책, 그리고 가치 배제의 이상』(Douglas 2009)에도 담겼다. 1절에서 거론했다시피, 해당 논문의 주요 내용은 과학자는 자신의 연구가 불러올 수 있는 의도하지 않은 결과를 고려할 도덕적 책임이 있으며, 그 결과를 어느 정도까지 예견하는 것이 타당한지에 대한 기준은 그가 속한 과학 공동체의 기준을 따른다는 것이었다. 또한 이를 통해 더글러스는 과학자의 책임과 자율성이 적대적인 관계에 있는 것이 아니라고 강조했다.

더글러스의 구체적인 주장은 다음과 같다. 더글러스는 우선 모든 합당한 도덕적 행위자가 지니는 "일반 책임(general responsibility)"을 논한 후, 과학자가 과학자로서 지니는 "역할 책임(role responsibility)"이 일반 책임을 면제해 주는지 검토한다. 먼저, 모든 인간은 자신의 행위가 원인이 된 결과에 대하여, 자신이 그 결과를 의도했을 때뿐만 아니라 의도하지 않았을 때도 어느 정도는 도덕적 책임이 있을 수 있다. 이는 인간으로 지는 일반 책임에 해당한다. 의도하지 않은 결과에 대해 도덕적으로 비난받을 수 있는 경우는 크게 두 가지가 있는데, 무모(reckless)했거나 과실(negligent)이 있었을 때다.[7] 여기서 더글러스는 법철학자 파인버그(Feinberg 1970)의 정의를 따라 합당하지 않은 위험을 알고서(knowingly) 초래한 것은 무모함이라 부르며, 모르고(unknowingly) 초래한 것은 과실이라 칭한다. 예를 들어, 목숨이 위태로운 환자를 병원에 이송하기 위해 과속 운전을 하는 것은 합당한 위험으로 볼 수 있으므로 무모함이나 과실에 해당하지 않는다. 그러나 단순히 재미를 위해 과속한다면 무모한 것이며, 얼마나 빠른 속도로 운전하고 있는지 확인하지 않는다면 과실이 있는 것이라 할 수 있다. 이 경우 운전자는 교통사고를 낼 의도가 없더라도 상당한 책임을 져야 할 것이다.

더글러스는 이러한 일반 책임을 과학자에게도 적용하며, 과학자 역

7) 이 논문에서는 영미법학의 번역을 따라 "reckless"와 "negligence"을 각각 "무모함"과 "과실"로 옮겼다. 번역에 대한 제안을 주신 편집인과 심사위원께 감사드린다.

시 자신의 행위가 불러올 수 있는 의도하지 않은 결과에 대해 무모하거나 과실을 저지르지 않을 도덕적 책임이 있다고 주장한다. 즉 과학자 또한 합당하지 않은 위험을 무릅쓰거나, 그러한 위험을 예견하고 완화하려는 노력을 게을리해서는 안 된다는 것이다. 이 주장은 더글러스의 유명한 "귀납적 위험으로부터의 논증(Argument from Inductive Risk; AIR)"과도 긴밀히 연결된다. 앞서 더글러스(Douglas 2000)는 러드너(Rudner 1953) 등의 주장을 기반으로 AIR을 발전시킨 바 있다. AIR의 핵심은 귀납적 추론을 할 때 오류의 위험, 즉 참인 가설을 거짓으로 잘못 판단하거나 거짓인 가설을 참으로 잘못 판단할 위험을 완전히 피할 수는 없으므로, 귀납적 추론에는 각 오류가 불러올 결과의 경중을 판단하는 비인식적(non-epistemic) 가치가 정당하게 개입한다는 것이다. 이를 과학자의 도덕적 책임에 대한 더글러스의 주장과 결합하면 다음과 같이 말할 수 있다. 과학자는 과학적 주장을 할 때, 본인의 주장이 틀렸을 경우 발생할 수 있는 사회적 파장을 예견하고 심사숙고해야 한다.

이처럼 과학자의 일반 책임을 논한 후, 더글러스는 과학자의 "역할 책임"이 "일반 책임"을 면제하는지 검토한다. 직업인 대부분이 그러하듯, 과학자 또한 인간으로서 지는 일반 책임 외에도 과학자라는 특정 역할을 맡음으로써 지게 되는 별도의 역할 책임이 있다. 연구 결과를 주의 깊게 보고할 의무 등이 여기에 속한다. 그런데 더글러스가 보기에 역할 책임이 일반 책임을 면제할 수 있는 경우는 오직 그 역할과 주요한 책임을 모두 명료하게 명시한 엄격한 체계가 갖춰져 있을 때뿐이다.[8] 예를 들어 변호사의 책임을 생각해 보자. 변호사가 아닌 일반 시민에게는 누군가의 범죄행위를 알게 되었을 때 이를 경찰서 등에 신고할 책임이 있지만, 변호사는 의뢰인의 범죄행위를 신고할 책임으로부터 면제받는다. 변호사에게는 변호사로서 의뢰인의 비밀을 지킬 역할 책임이 있으며, 이 역할 책임은 관련된 일반 책임으로부터 면제를 주기 때문이다. 하지만 더글러스는 이는 형사 사법 체계가 엄격하게 갖춰져 있기에 가능한 일일 뿐, 과학에는 이러한 체계가 없다고 지적한다.

관련하여, 과학자는 역할 책임만 진 채로 과학 연구에 몰두하게 하

8) Douglas (2009), p. 73.

고, 그 외의 일반 책임은 다른 사람에게 전가하는 방안도 생각해 볼 수 있다. 그러나 더글러스는 과학자의 일반 책임을 다른 사람에게 전가하는 것은 가능하지도, 바람직하지도 않다고 주장한다. 과학 연구가 불러올 잠재적 결과를 가장 잘 예견하고 고려할 수 있는 사람은 바로 그 연구를 수행하고 있는 과학자 자신이다. 이러한 책임을 다른 누군가가 완전히 대신할 수 있으려면 연구의 모든 세부 사안을 지속해서 파악하고, 필요한 경우 연구를 변경하거나 멈추게 할 권한을 가져야 한다. 그런데 이는 곧 과학자가 그들이 가진 자율성을 대부분 포기해야 함을 의미한다.[9] 다시 말해, 20세기 말까지의 지배적 견해와 다르게, 더글러스는 과학자가 스스로 도덕적 책임을 져야 자율성도 가질 수 있다고 강조한 것이다.

그렇다면 과학자는 **어느 정도까지** 자신의 연구가 불러올 잠재적 결과를 예견해야 하는가? 이 지점이 바로 이 논문에서 내가 주목하는 부분이자, 과학 공동체에 대한 더글러스의 견해가 잘 드러나는 대목이다. 우리는 어떤 사람이 범죄를 저질렀을 때 대개 그 사람의 먼 조상을 도덕적으로 비난하지는 않는다. 조상의 존재로부터 후손의 범죄까지 이어지는 인과 사슬이 존재한다고 볼 수 있으나, 먼 후손의 행위를 모두 예견하기를 기대하는 것은 우리 사회의 통념상 합당하지 않기 때문이다. 과학자도 마찬가지다. 자신의 연구가 불러올 모든 잠재적 결과를 완벽하게 예견할 필요는 없고, 그럴 수도 없다. 다만 합당히 예견할 수 있는 결과에 대해서는 예견해야 한다. 그리고 더글러스의 주장은 과학자가 어느 정도까지 예견하기를 요구하는 것이 합당한가에 대한 기준은 그가 속한 과학 공동체가 제공한다는 것이다. 나는 이를 더글러스의 "공동체 기준"이라 부르겠다.

더글러스는 과학 공동체나 공동체 기준이 무엇인지에 대해 구체적인 설명이나 엄격한 논증을 제시하지 않는다. 그럼에도 더글러스의 논의를 따라가 보면 그가 과학 공동체는 **단일한** 기준을 **신속하게** 제공할 것이라는 가정을 바탕에 두고 있음을 알 수 있다. 우선, 더글러스는 공동체에 대해 논할 때 법학에서의 합당한 예견에 대한 기준에 바탕을 두고

9) *Ibid*., pp. 74-5.

있다. 미국 법체계에서는 피고인의 과실을 판단할 때 "합당한 사람"이라면 어느 정도까지 예견하고 주의를 기울였을지에 대한 질문을 중요하게 다룬다. 이때 합당한 사람이란 대개 "공동체" 내의 일반적인, 혹은 이상적인 사람으로 간주한다. 이에 대한 구체적인 해석은 학자마다 다르지만, 공동체 내의 다수결 혹은 이상적 규범에 따라 정해지는 단일한 기준이 존재한다는 가정이 미국 법체계에 자리하고 있다고 볼 수 있다.[10]

더글러스 역시 과학자의 합당한 예견에 대해 논하면서 이러한 공동체의 기능을 강조한다. 단, 더글러스는 과학자의 예견에 대한 기준은 일반적인 시민 공동체보다는 과학 공동체의 기준을 따라야 한다고 말한다. 더글러스의 아래 서술을 보자.

> 과학자에게 예견할 수 없는 결과에 대한 책임을 묻는 것은 부당하다. 과학자가 누구든 공유할 기본적인 고려와 예견의 기준을 충족하도록 기대하는 것이 합당하며, 이러한 예견에 대한 합당한 기대는 과학 공동체 내 동료들과의 비교를 통해 판단되어야 한다.[11]

과학자 누구든 공유할 예견의 기준이 있을 것이라는 더글러스의 서술로부터 우리는, 어떤 사안에 대해 모두에게 받아들여질 단일한 기준이 있을 것이라는 그의 기대를 읽을 수 있다. 더글러스는 과학 공동체에 이러한 단일한 기준을 기대하는 것이 일반 시민 공동체와 비교했을 때 더 무리한 요구가 아니며, 오히려 과학 공동체는 이러한 기준을 더 손쉽게 형성할 수 있다고 여긴다. 이는 끊임없이 소통하는 과학 공동체의 특성 때문이다. 아래 서술을 보자.

> 과학자들은 다른 과학자들과 거의 끊임없이 소통하고 경쟁하는 공동체에서 일하기 때문에 예견 가능한 것과 불가능한 것을 쉽게 판단할 수 있다. 과학의 다른 아이디어와 마찬가지로 오류의 잠재적인 결과는 빠르게 확산하며 과학자들은 함정과 위험, 불확실성에 대해 쉽게

10) 미국 법체계에서의 합당한 사람 개념에 대한 자세한 논의는 Scalet (2003) 및 Lefevere and Schliesser (2014) 참조.

11) Douglas, *op. cit.*, p. 84.

논의한다.[12]

다시 말해, 더글러스는 끊임없이 소통하는 과학 공동체의 특성상 단일한 기준이 신속하게 생성되리라 가정하고 있는 것이다. 과학 공동체에 대한 더글러스의 이러한 가정을 두고, 르페버러와 슐리서(Lefevere and Schliesser 2014)는 마치 경제학에서 말하는 "효율적 시장(efficient market)"의 과학 버전 같다고 비유하기도 했다. 정보의 공유와 의사소통에 아무런 장애물이 존재하지 않으며, 새로운 정보가 가격과 같은 특정 기준에 즉각적으로 반영되는 공동체 말이다.

공동체 기준이 어떻게 생성되고 작동하는지를 설명하는 예시 중 하나로 더글러스는 1970년대 유전자 재조합 기술을 둘러싼 생물학 공동체의 사례를 제시한다. 당시 생물학자들 사이에서는 유전자 재조합 기술을 연구하는 과정에서 의도치 않게 생물학적 위해 물질이 생성되어 실험실 밖으로 방출될 가능성에 대한 우려가 빠르게 확산했다. 이에 따라 이들은 1974년에 모라토리엄(moratorium), 즉 자발적으로 연구를 잠시 중단할 것을 선언했다. 다시 말해, 생물학 공동체 내에서는 유전자 재조합 기술에 대한 정보 공유와 토론이 신속하게 이루어졌으며, 위해 물질이 의도치 않게 생성되고 방출될 수 있다고 예견하는 것이 당시 생물학 공동체의 기준에서 합당한 것으로 받아들여졌다고 볼 수 있다.[13]

이상 과학자의 합당한 예견에 대한 더글러스의 주장과 그 바탕이 되는 "공동체 기준"에 대한 그의 가정을 살펴보았다. 내가 이 논문에서 다루고자 하는 질문은 더글러스가 20여 년 전에 가정한 이러한 "공동체 기준"이 오늘날의 과학, 특히 인공지능 기반 과학에서 얼마나 유효한가 하는 점이다. 더글러스가 서술한 과학 공동체의 모습을 오늘날의 인공지능 기반 과학에서도 기대할 수 있는가? 나는 과학자에게는 의도하지 않은 결과를 고려할 도덕적 책임이 있다는 더글러스의 주장에는 대체로 동의하지만, 더글러스가 말한 공동체 기준은 수정이 필요하다고

12) *Ibid.*, p. 83.

13) 유전자 재조합 기술을 둘러싼 생물학자들의 대응에 대한 비판적 논의는 우태민, 신유정, 박범순 (2025) 참조.

주장하고자 한다. 이제 다음 절에서는 인공지능 기반 과학 연구의 특징을 살펴보며 더글러스의 공동체 기준의 문제점을 지적하겠다.

3. 인공지능 기반 과학 연구: 공동체 기준의 문제점

오늘날의 인공지능 기반 과학 연구는 더글러스가 상정한 과학 연구와는 사뭇 다른 특징을 지니고 있다. 이 절에서는 더글러스의 공동체 기준의 한계를 드러내는 인공지능 기반 과학 연구의 특징 세 가지를 소개하며 이를 각각 다학제의 문제, 느린 기준 설정의 문제, 편향된 공동체의 문제라 명명하겠다. 명확하게 말하자면, 이 세 가지 특징은 인공지능 기반 과학 연구에만 한정되는 특징은 아니지만, 과학 전체로 일반화할 수 있는 특징이라 보기는 힘들다.[14] 이 절의 목표는 인공지능 기반 연구는 더글러스의 주장을 만족하기 어려움을 보이는 것이며, 과학 전체로 일반화하여 더글러스의 주장을 전면적으로 반박하려는 것은 아님을 밝혀두겠다.

3.1. 다학제의 문제

인공지능 기반 과학 연구에서 눈여겨볼 첫 번째 지점은 인공지능 연구자와 타 영역 전문가(domain expert), 특히 실험가의 견해 차이에 대한 문제다. 널리 알려졌다시피 인공지능 모델의 성능은 데이터의 양과 질에 크게 의존한다(Mock et al. 2023; Nature 2023). 오늘날 인공지능의 핵심이라 할 수 있는 기계학습은 방대한 데이터를 컴퓨터가 스스로 학습하고 일반화하게끔 하는 기술을 일컫는다. 편향되었거나 오류가 많은 데이터를 토대로 학습한 모델은 그만큼 부정확한 예측을 할 수 있으므로 데이터의 양과 질을 관리하는 것은 인공지능 기반 과학 연구의 핵심 요소로 꼽힌다. 이러한 높은 데이터 의존성을 완화하고자 최근에

14) 구체적으로는, 3.3절의 핵심 내용은 일반화가 가능하겠지만, 3.1절과 3.2절에서 다룰 내용은 사례별로 검토해 보아야 할 것이다.

는 실험실 등에서 얻은 데이터 대신 인공지능 모델이 생성한 결과를 이용해 모델을 재귀적으로 학습시키려는 시도도 있었다. 그러나 이러한 재귀적 학습은 인공지능 모델의 "붕괴"를 불러와 매우 오류가 많고 기이한 결과 생성으로 이어진다는 연구가 발표되기도 했다(Shumailov et al. 2024).

이와 같은 인공지능 모델의 높은 데이터 의존성으로 인하여 인공지능 연구자와 실험가의 긴밀한 학제 간 협력은 인공지능 기반 과학 연구의 성공을 좌우할 요인 중 하나로 지목되기도 한다. 예를 들어 신약 개발 연구에서는 의약화학자 등과의 협업 없이는 좋은 인공지능 모델을 만들기 어렵다(Griffen et al. 2020). 이들 실험가가 화학물질 합성 등을 통하여 양질의 실험 데이터를 제공해 주어야 인공지능 모델을 훈련시킬 수 있으며, 모델이 올바르게 훈련되었는지 확인할 때도 모델이 출력한 결과를 실험 데이터와 비교해야 하기 때문이다.[15]

그런데 문제는 각 과학 분과는 영역 특이성(domain specificity)을 지닌다는 데 있다. 영역 특이성과 학제 간 협력을 다룬 대표적인 과학철학 연구로는 맥레오드와 너세시안의 민족지 연구가 있다(MacLeod and Nersessian 2013, 2014; MacLeod 2018). 영역 특이성은 각 분야에서 다루는 주제나 문제 유형의 범위가 좁고, 해당 주제만을 효과적으로 다루기 위해 각 분야의 체계가 세밀하게 조정되어 있어 유연성이 부족한 것으로 특징지을 수 있다.[16] 맥레오드는 이러한 영역 특이성이 계산과학자와 실험가 사이의 학제 간 협업을 방해하는 과정을 분석한다. 예컨대 실험에 무지한 계산과학자는 효과적인 시뮬레이션 모델 구축을 위해 실제로 실험실에서 생성하기에는 매우 수고로운 특정 데이터를 제공할 것을 실험가에게 요청할 수 있다. 이때 계산과학에 무지한 실험가는 그러한 수고로움을 감수할 만한 이유를 찾지 못해 해당 데이터를 생성하는 것에 거부감을 느끼게 되고, 이는 계산과학자와 실험가 사이의 관계 악화 및 소통 단절로 이어질 수 있다.[17] 계산과학자와 실험가

15) 인공지능 모델이 올바르게 훈련되었는지 확인하는 과정을 모델 검증(validation)이라 부르기도 한다.
16) MacLeod (2018), p. 703.
17) *Ibid*., pp. 707-8.

의 이러한 영역 특이성에 대한 분석은 인공지능 기반 과학 연구를 이해하는 데도 유용하다. 보다 최근에는 인공지능 기반 과학 연구에서 자연과학자가 인공지능 사용을 꺼리는 이유에 대한 민족지 연구도 이루어졌는데, 마찬가지로 자연과학자들과 인공지능 연구자 사이의 서로 다른 배경지식 등이 강조되었다(Simkute et al. 2024).

이와 같은 인공지능 연구자와 실험가의 영역 특이성 문제는 더글러스의 공동체 기준에 의문을 제기한다. 2절에서 언급했듯이, 더글러스의 주장은 과학 공동체가 합당한 예견에 대한 단일한 기준을 제시할 수 있다는 기대를 바탕으로 한다. 그러나 이러한 기대와는 달리, 인공지능 기반 과학 연구에서는 의도하지 않은 결과를 어디까지 예견하는 것이 합당한 것인지에 관해 인공지능 연구자 공동체와 실험가 공동체가 서로 다른 기준을 제시할 수 있다. 즉, 같은 사안을 두고서도 한 가지가 아닌 두 가지 또는 그 이상의 기준이 존재할 수 있다는 것이다.

구체적인 예시는 다음과 같다. 몇 년 전 『네이처 기계 지능(Nature Machine Intelligence)』에 신약 개발을 위해 만들어진 인공지능 모델이 독성 물질 개발에 이용될 수 있다는 연구 결과가 발표되었다(Urbina et al. 2022). 해당 연구에 의하면 본래 신약 후보 물질을 찾기 위해 만들어진 기존 인공지능 모델을 독성 물질을 탐색하도록 소폭 수정하였더니 6시간 만에 약 4만 종의 독성 물질을 제시했다고 한다. 이러한 연구 결과에 따라, 인공지능 연구자 공동체는 신약 개발을 위해 만들어진 인공지능 모델이 독성 물질 개발에 이용되지 않도록 각별한 주의를 기울이도록 요구할 수 있을 것이다. 그러나 미국 화학공학회의 『화학·공학 뉴스(Chemical and Engineering News; C&EN)』에 실린 인공지능 기반 연구 관련 기사(Mullin 2023)에서는 다소 다른 견해가 제시된다. 해당 기사의 인터뷰에 따르면 안전한 분자를 설계하기 위해 만들어진 인공지능 모델은 독성 분자를 설계하는 모델로 쉽게 전환될 수 있지만, 제안된 분자 구조에 따라 실험실에서 실제 물질을 합성하는 것은 높은 전문성이 필요한 어려운 일이다. 즉, 독성 물질을 쉽게 설계할 수 있다고 해서 실제로 쉽게 만들 수 있는 것은 아니라는 뜻이다. 이러한 실험의 어려움으로 미루어 보아, 신약 개발을 위한 인공지능 모델이 의도치

않게 독성 물질 개발로 이어질 위험에 대해 화학 실험가 공동체는 인공지능 연구자 공동체와는 다른 기준을 가질 것으로 짐작할 수 있다.

3.2. 느린 기준 설정의 문제

두 번째로 지적할 문제는 인공지능이 불러올 수 있는 잠재적 위험에 대한 공동체의 합의가 이뤄지는 속도보다 인공지능 기술의 발전 및 응용 속도가 지나치게 빠르다는 점이다. 최근 인공지능의 급격한 발전 속도에 대한 경각심을 요구하는 기사와 보고서가 쏟아지고 있다(Maslej et al. 2024; Bengio et al. 2024). 인공지능 기술의 발전 속도가 기존의 평가 및 측정 기준을 무력화할 정도로 빠르므로 이러한 기준을 재설계해야 한다는 의견도 곳곳에서 제기되고 있다(Criddle 2024). 한편, 더글러스의 공동체 기준은 연구에 관한 정보와 우려가 실시간으로 공유되며, 합당한 예견에 대한 기준을 **신속하게** 제공하는 과학 공동체를 가정한다. 하지만 인공지능 기술의 급속한 발전 속도를 고려한다면, 인공지능 연구자 공동체(혹은 다른 어떤 과학 공동체라도)가 인공지능이 불러올 잠재적 위협에 관한 토론을 현황에 맞게 진행하고 이에 대한 기준을 제때 제공할 수 있을지 의심스럽다.

기술의 발전 속도가 문제라면, 1970년대에 유전자 재조합 기술 연구를 자발적으로 잠정 중단했던 생물학자들의 선례를 따라, 공동체가 적절한 기준을 마련할 수 있을 때까지 인공지능 개발을 중단하자는 의견이 있을 수 있다. 1절에서 언급한 거대 인공지능 연구를 잠시 중단하자는 삶의 미래 연구소의 공개서한이 그 예시에 해당한다(Future of Life Institute 2023). 그러나 인공지능의 경우 잠정적 연구 중단이 실현될 가능성은 희박해 보인다. 유전자 재조합 기술의 경우 특정한 고가의 실험실 자원과 상당한 전문성 및 지식이 필요했기 때문에 과거 이 기술을 연구할 수 있었던 사람은 상당히 제한적이었다. 반면, 오늘날 인공지능 기술에는 상대적으로 훨씬 많은 사람이 쉽게 접근할 수 있다. 따라서 인공지능 연구자 다수가 연구 중단을 제안하더라도, 전 세계 모든 사람의 연구를 막는 것은 불가능에 가깝다. 실제로 삶의 미래 연구소의 공개서한에는 인공지능 업계의 권위자 다수를 포함한 3만 명 이상의 개

인이 서명하였지만, 공개서한에서 제시된 기한에 거대 인공지능 연구가 중단되기는커녕 많은 인공지능 기업은 더욱 거대한 인공지능 모델을 훈련시켰다(Aguirre 2024).

3.3. 편향된 공동체의 문제

마지막으로 강조할 문제는 합당한 예견의 기준을 제시해야 하는 공동체 자체가 편향될 수 있다는 점이다. 이는 더글러스의 주장에 대한 르페버러와 슐리서의 비판에서 잘 드러난다(Lefevere and Schliesser 2014). 더글러스의 주장에 따르면, 개별 과학자가 공동체 기준에 맞게 행동했다면 그는 어떠한 도덕적 비난도 받지 않아야 할 것이다. 그런데 만약 공동체 기준 자체에 문제가 있었고, 그 기준을 따른 결과 개별 과학자가 매우 무모하거나 과실이 있는 행위를 했다면, 그는 도덕적으로 책임 있게 행동했다고 봐야 하는가? 이러한 질문을 통해 르페버러와 슐리서는 과학 공동체를 과학자의 도덕적 책임에 대한 기준으로 삼는 더글러스의 주장은 과학 공동체의 특성 자체가 도덕적 평가의 대상이 될 수 있음을 간과한다고 지적한다.

편향의 문제는 현대 인공지능 윤리의 대표적인 주제 중 하나다(Fazelpour and Danks 2021; Angwin et al. 2022). 현실의 데이터를 학습하는 인공지능 모델이 현실 세계의 성차별, 인종차별 등을 답습하고 강화할 수 있다는 우려가 논의의 주를 이룬다. 만약 데이터를 수집하고 활용하는 인공지능 연구자가 편향과 차별 문제에 대해 적극적인 관심을 가진다면, 학습 데이터에 내재한 문제를 어느 정도 개선할 수도 있을 것이다. 그러나 핵심은 인공지능 연구자 공동체 자체도 편향되어 있을 가능성이 있다는 데 있다(Kuhlman et al. 2020; Young et al. 2023).[18] 이러한 연구자 공동체의 편향은 차별받는 집단에 특히 악영향을 줄 수 있는 의도치 않은 결과를 예견하는 데 있어 과실을 불러올 수 있다. 실제로, 범죄자의 재범 위험을 예측하는 알고리즘인 콤파스

18) 인공지능 외 분야 과학 공동체의 편향과 과학적 지식 생산에 대한 분석은 Longino (1990), Lloyd (2005) 등을 참조. 4절에서 이에 대한 추가적인 논의를 제공한다.

(COMPAS)의 예측 결과를 분석한 결과, 재범을 저지르지 않은 흑인을 재범 고위험군으로 잘못 예측한 비율이 백인을 재범 고위험군으로 잘못 예측한 비율보다 두 배 가까이 높다는 보고가 있었다(Angwin et al. 2016).

아울러, 인공지능 연구자 공동체와 IT 기업 간의 관계 또한 주목할 만하다. 오늘날 인공지능 연구는 강력한 자금력을 보유한 IT 기업 및 해당 기업 소속 연구원들에 의해 주도되고 있다. 또한 구글이나 마이크로소프트 같은 거대 IT 기업이 영향력 있는 인공지능 윤리 원칙 및 지침을 발표해 왔다는 점에서 알 수 있듯이, IT 기업은 인공지능 윤리 담론 형성에도 상당한 영향력을 행사해 왔다(Hagendorff 2020). 최근 여러 과학철학자는 사적 이익을 추구하는 기업이 데이터의 생성 및 수집, 연구 설계, 논문 출판 등에 영향을 미침으로써 과학 연구에 인식적 편향을 초래할 수 있다고 주장한 바 있다(Elliott and McKaughan 2009; Holman and Elliott 2018). 이러한 맥락에서, 인공지능 연구 역시 합당한 예견에 대한 기준이 기업의 이익에 부합하는 방향으로 편향되어 있지는 않은지 의심해 볼 수 있다.[19]

4. 제안: 맥락과 다원주의

위에서 지적한 더글러스의 “공동체 기준”의 문제점을 해결하기 위해, 이 절에서는 인식적 기준의 맥락 의존성과 다원주의를 바탕으로 합당한 예견의 기준을 수정할 것을 제안한다. 내 제안의 핵심은 첫째, 의도하지 않은 결과에 대한 합당한 예견의 기준은 특정 연구 프로젝트의 맥락을 중심으로 설정되어야 하며, 둘째, 과학 공동체는 단일한 기준을 제시하는 대신 다양한 관점을 적극적으로 고려해야 한다는 것이다. 이때 “맥락 의존성”과 “다원주의”라는 용어는 철학에서 다양한 의미로 사용될 수 있지만, 이 논문에서는 과학과 비인식적 가치에 대한 과학철

19) 인공지능 윤리와 기업의 영향에 대한 자세한 논의는 한혜정 (2021), Ochigame (2022) 등을 참조.

학 및 사회인식론 논의에서의 용어 사용을 주로 참조할 것이다.

4.1. 연구 프로젝트와 맥락 의존성

먼저 첫 번째로 제안할 점은 "합당함(reasonableness)"의 의미와 깊은 관련이 있다. 2절에서 언급했듯이 더글러스는 합당한 예견이 무엇인지에 대해 설명할 때 법철학에서 논의를 빌려온다. 미국의 법체계에서 "합당한 사람(reasonable person)"이라는 개념은 중심적인 역할을 차지한다. 여기서 "합당한 사람"은 주로 해당 공동체에서의 "일반적인 사람(ordinary person)" 혹은 "이상적인 사람(ideal person)"으로 해석된다. 그런데 법철학자 스칼릿(Scalet 2003)은 이러한 해석이 실제 상황에서 사용할 수 있는 안정적인 지침을 제공할 수 없다고 비판하며 합당함에 대한 "쌍안경 관점(Binocular View)"을 제시한다. 일반적인 사람 혹은 이상적인 사람이라는 추상적인 기준에 전적으로 기대는 대신, 첫째, 판단하고자 하는 사례에서의 중요한 특성을 식별하고, 둘째, 해당 특성을 가진 사람의 실제 행동이나 신념을 추정하자는 것이다.

스칼릿의 개별 사례 및 특성에 대한 강조는 과학과 가치 논의에서 제시된 인식적 기준의 맥락 의존성과 연결될 수 있다. 대표적으로 헬렌 론지노(Longino 1990; Longino 2002)의 맥락적 경험주의(contextual empiricism)를 들 수 있다. 론지노는 어떤 데이터가 가설을 지지하거나 반박하는지, 즉 해당 가설에 대한 증거로 받아들여질지는 데이터 자체로 결정되는 것이 아니라, 과학 연구의 배경 가정에 의존한다고 주장한다. 또한 이러한 배경 가정은 해당 과학적 추론이 이루어지는 맥락에서 비인식적 가치에 영향을 받는다고 본다. 다시 말해, 데이터의 증거력은 개별 과학적 추론이 이루어지는 구체적인 맥락에 영향을 받는다는 것이다.[20] 여기서 론지노는 더글러스와 유사하게 인식적 기준의 구심점으로서의 "공동체"의 역할에 주목한다. 그러나 론지노가 서술하고 있는

20) 론지노의 주장을 뒷받침하는 사례 연구는 페미니스트 과학철학에서 다수 제시되었다. 예를 들어, Lloyd (2005) 참조. 인간 진화 연구에서 오르가슴의 역할을 분석할 때 과학계의 남성중심적인 편향이 여성의 오르가슴에 대한 데이터를 간과하였다고 주장한다.

공동체는 더글러스가 상정한 것보다는 훨씬 유연하고 가변적이다. 론지노는 증거력의 기준은 특정 연구 맥락에서의 일시적인 목표나 가치 등에 의해 비판받고 변화할 수 있다고 언급한다. 더글러스가 추상적이고 단일한 공동체 기준을 강조한다면, 론지노는 개별 과학적 추론이 이루어지는 구체적인 맥락에 따라 증거의 기준이 어떻게 변화하는지에 더욱 관심을 쏟는 셈이다.

과학 연구의 맥락 의존성을 강조한 또 다른 저명한 과학철학자로는 필립 키처(Kitcher 2001, 2011)가 있다. 그는 과학 연구가 사회적 가치를 반영하여 민주적으로 진행될 필요가 있다고 주장하는데, 그가 서술한 이상적인 과학의 특징에는 "과학적 중요성(scientific significance)"에 따라 과학 연구의 의제가 정해지고 예산이 분배되는 것이 포함된다.[21] 여기서 핵심은 이 과학적 중요성이 자연에 대한 근본적 이해나 실용적 문제 해결 중 한 가지 추상적 기준에 따라 전적으로 정해지는 것이 아니라, 두 가지 목표가 얽히고설킨 역사적 산물로서 결정된다는 것이다. 키처는 복제 양 "돌리(Dolly)" 탄생 프로젝트를 예시로 든다.[22] 그가 보기에 복제 양 탄생이라는 특정한 연구 프로젝트가 과학적으로 중요했던 이유는 생명체의 발달에 대한 이해라는 인식적 목표와 가축 개량이라는 실용적 목표가 서로 복잡하게 얽혀 있었기 때문이다. 이는 단순히 두 종류의 목표가 모두 중요했다는 의미가 아니다. 과거의 사회적·실용적 조건이 오늘날의 자연 탐구를 돕고, 반대로 과거의 자연 탐구가 오늘날의 실용적 목표 달성을 돕듯이, 인식적 중요성과 실용적 중요성은 완전히 분리될 수 없다는 의미다. 이러한 키처의 분석은 특정 연구의 과학적 중요성은 추상적이고 단일한 기준으로 결정되는 것이 아닌 그 연구가 수행될 구체적인 맥락에 의존함을 보여준다.

증거의 기준에 대한 론지노의 맥락 의존성, 그리고 과학적 중요성에 대한 키처의 맥락 의존성에 덧붙여, 나는 "합당한 예견"에 대한 기준을

21) 키처는 그의 구상을 "질서 정연한 과학에 대한 이상(ideal of well-ordered science)"이라 명명한다. 키처의 이상에서는 의제 선정 및 예산 배정, 연구의 수행, 연구의 활용 등이 민주적 숙고를 통해 진행된다.

22) Kitcher (2001), pp. 63-82.

정할 때도 역시 특정 연구 프로젝트가 진행되는 구체적인 맥락에 주목할 것을 제안한다. 즉, 어디까지 예견하고 주의를 기울이는 것이 합당한지에 대한 기준을 정할 때 연구 프로젝트의 구체적인 목표, 프로젝트의 결론이 제시되어야 하는 기간, 프로젝트에서 활용할 수 있는 기술이나 지식 등을 세부적으로 고려해야 한다는 뜻이다. 내가 여기서 말하는 연구 프로젝트란 특정하고 구체적인 목적을 달성하기 위하여 주로 소규모 과학자 집단이 진행하는 과업을 일컫는다. 한 가지 연구 프로젝트에는 다양한 분과의 과학자가 함께 참여할 수도 있다. 위에서 언급한 복제 양 돌리 탄생 프로젝트나 2024년 노벨화학상 수상으로 이어진 단백질 구조 예측 인공지능 개발 프로젝트 등이 그 예시이다.

이처럼 추상적이고 단일한 과학 공동체라는 이상에 전적으로 기대는 대신 특정 연구 프로젝트의 맥락에 주목한다면, 우리는 절대적인 진리 탐구나 인류 전체의 번창이라는 추상적인 목표만을 좇는 것에서 나아가 지엽적이고 단기적인 차원에서 연구 목표를 구체화하고 연구의 쓰임새 등을 효과적으로 고려할 수 있게 된다. 아울러, 한 분과의 과학자 전체가 숙지하기는 힘든 특정 연구 프로젝트와 관련된 세부적인 지식이나 기술 등을 자세히 검토할 수 있게 된다. 이러한 이점을 활용하면 3절에서 논의한 다학제의 문제와 느린 기준 설정의 문제를 해결할 실마리를 찾을 수 있다.

구체적으로 3절에서 예시로 들었던 분자 설계 인공지능 모델을 생각해 보자. 신약 개발에 활용하기 위하여 분자 설계 인공지능 모델을 개발할 때, 화학 합성에 대해 무지한 인물로만 구성된 인공지능 연구자 공동체라면 그들이 개발한 인공지능 모델로 인해 독성 물질 개발이 매우 쉬워질 것이라 짐작할 수 있다. 그러나 이 연구 프로젝트의 구체적인 맥락, 즉 화학 합성과 신약 개발에 대한 세부적인 지식 등을 함께 고려한다면 분자 설계 인공지능이 악용될 가능성에 대해 더 타당하고 구체적인 예견을 할 수 있을 것이다. 즉, 한 공동체 내의 한 가지 추상적인 기준에 전적으로 기대는 것이 아니라, 연구 프로젝트의 맥락에 맞는 세부 지식 등을 반영하여 구체적인 기준을 세움으로써 다학제의 문제를 해결할 수 있는 것이다. 마찬가지로 3절에서 지적한 느린 기준 설

정의 문제의 경우, 우리의 초점을 공동체에서 특정 연구 프로젝트로 옮긴다면 이는 더 이상 심각한 문제가 되지 않는다. 공동체가 절대적인 기준을 제시해야 하는 것이 아니라, 해당 맥락에 따라 구체적인 기준을 정하면 되기 때문이다.

명확히 말하자면, 연구 프로젝트의 구체적인 맥락에 주목하자는 나의 제안은 공동체의 역할 전체를 부정하는 것은 아니다. 공동체 내에서의 다수의 의견 등은 주요 참조가 될 수 있을 것이다. 그러나 한 종류의 공동체 기준이 절대적인 역할을 하지 않으며, 또한 실제로 적용할 기준을 구체화하기 위해서는 연구 프로젝트의 맥락에 대한 고려가 필요하다는 주장이다. 이러한 기준을 만족하는지에 대해서는 다양한 주체가 평가할 수 있는데, 단 해당 맥락을 이해하는 주체여야 할 것이다.

이때, 평가의 기준이 프로젝트의 맥락에 따라 정해지고, 평가받는 대상은 해당 프로젝트 내의 과학자라면, 평가 대상과 평가 기준이 같아지는 것이 아니냐는 우려가 제기될 수 있다. 그런데 이러한 우려는 더글러스의 주장에도 적용할 수 있다. 다시 말해, 더글러스의 주장에서도 역시 평가 기준을 과학 공동체가 제공하고, 평가받는 대상은 해당 과학 공동체의 구성원이므로, 비슷한 우려가 제기될 수 있는 것이다. 이러한 문제에 대하여 내가 다음 절에서 제시할 다원주의는 해결의 실마리를 제공할 수 있다.[23] 이제 다음 절에서는 이 문제를 구체적으로 다루어 보겠다.

4.2. 동원의 책임과 다원주의적 공동체

위에서 설명한 첫 번째 제안에 대하여, 다학제 연구 프로젝트라 하더라도 그 프로젝트에 참여하는 과학자 개개인은 자신의 전공 분야의 공동체 기준에 기대지 않겠냐는 의견이 제기될 수 있다. 즉, 연구 프로젝트의 구체적인 맥락에 주목하자는 제안은 결국 더글러스의 공동체 기준과 양립할 수 있는 형태로 이를 보완하는 역할을 하는 것이지, 공동체 기준을 반박하지는 않는다는 의견이다. 그러나 지금부터 설명할 두

23) 이 지점을 명확히 하도록 도움을 주신 심사위원께 감사를 전한다.

번째 제안은 더글러스의 공동체 기준에 대한 적극적인 수정을 요청한다. 더글러스는 과학 공동체가 합당한 예견에 대한 단일한 기준을 제공할 것이라 가정하지만, 나는 과학 공동체는 단일한 기준을 제시하는 대신 다양한 관점을 적극적으로 고려해야 한다고 제안한다. 또한, 이러한 다원주의적 과학 공동체를 조성하기 위해 일부 개별 과학자에게 동원(mobilize)의 책임을 부여하자고 제안한다.

나의 제안은 르페버러와 슐리서(Lefevere and Schliesser 2014)의 인식적 다원주의에 기반을 두고 있다. 3절에서 언급한 것처럼, 르페버러와 슐리서는 더글러스의 주장이 과학 공동체 자체가 편향되어 있을 가능성을 반영하지 못한다고 비판한 바 있다. 이에 따라 그들은 과학 공동체가 집단적인 과실을 방지하기 위하여 다양한 관점을 적극적으로 고려해야 할 의무가 있다는 다원주의적 주장을 제시한다. 여기서 유의할 점이 두 가지 있다. 첫째, 르페버러와 슐리서의 인식적 다원주의는 개별 과학자가 아닌 과학 공동체에 초점이 맞춰져 있다. 이들의 주장은 더 랑허(De Langhe 2009)의 다원주의에 바탕을 두는데, 더 랑허는 다원주의를 "현실에 대한 다양한 가능한 관점의 타당성을 적극적으로 인정하는 인식적 입장"[24]으로 정의한다. 동시에 그는 자신의 견해는 개별 과학자가 본인의 연구를 수행할 때 다원주의를 받아들이도록 설득해야 한다는 것이 아니라, 과학 공동체의 역할에 집중하자는 것이라고 강조한다.[25] 그리고 둘째, 르페버러와 슐리서의 주장은 규범적이다. 즉, 더글러스의 공동체 기준은 공동체의 책임에 대한 규범적 진술이라기보다는 과학 공동체는 그러한 기준을 자연스럽게 제시할 것이라는 서술적 진술에 가깝지만, 르페버러와 슐리서는 과학 공동체는 다양한 관점을 적극적으로 고려하도록 노력해야 한다는 규범적 주문을 하고 있다.

어떻게 해야 이러한 다원주의적 과학 공동체를 조성할 수 있는지는 르페버러와 슐리서가 자세히 논하지 않는다. 이 방법에 대해서는 폴리티(Politi 2025)의 주장에서 단서를 얻을 수 있다. 폴리티는 우선 힌드릭스(Hindriks 2019) 및 플라이셔와 셰셸야(Fleisher and Šešelja 2023)

24) De Langhe (2009), p. 87.
25) *Ibid.*, p. 96.

의 논의를 참조하여, 집단적인 해악을 방지하기 위한 두 가지 단계의 책임을 설명한다. 첫 번째 단계는 개개인이 다른 사람을 "동원"할 책임이다. 그리고 그렇게 동원된 다수가 해악을 방지하는 것이 두 번째 단계이다. 예를 들어, 12명의 아이가 바다에 빠졌고, 주변에 어른 보호자 3명이 있다고 하자. 어른 개인이 헤엄을 쳐서 아이를 구한다면 각자 1명의 아이밖에 구할 수 없지만, 근처에 있는 보트를 이용한다면 12명의 아이를 모두 구할 수 있다. 그런데 보트를 움직이기 위해서는 최소 2명의 어른이 필요하다. 이러한 상황에서 우리의 직관은 어른 2명이 힘을 합쳐 보트를 움직여 12명의 아이 모두를 구하는 것이 가장 바람직하다는 것이다.[26] 이처럼 개인 혼자서는 대처하기 어렵지만 여럿이 힘을 합치면 해결할 수 있는 해악이 있을 때, 힘을 합치도록 다른 사람을 설득해야 한다는 것이 개인이 갖는 동원의 책임이다.

폴리티는 이러한 동원의 책임에 대한 논의를 과학 공동체의 도덕적 책임에 대한 논의로 확장한다. 단, 그는 모든 개별 과학자에게 그러한 책임을 부여하는 것은 바람직하지 않다고 강조한다.[27] 과학 공동체처럼 규모가 큰 공동체에서 개개인은 자신의 행동이 실질적인 영향을 미치지 않을 것으로 생각할 수 있으며, 이는 동원된 집단의 형성을 저해할 수 있기 때문이다. 이에 따라, 폴리티는 윤리적 노동의 분업을 제시한다. 즉, 과학자 공동체 내에서 인식적 노동의 분업이 필요한 것처럼 윤리적 노동 역시 분업이 필요하며, 지도적 위치에 있거나 윤리 관련 전문성을 지닌 개별 과학자 일부가 윤리적 노동을 담당하면 된다는 것이다. 이들 소수의 담당자가 누구를 어떻게 동원하고 업무를 조직할 것인지 결정하면, 공동체의 다른 구성원들은 이들 담당자가 제대로 업무를 수행하는지 감독할 수 있을 것이다.

이상 서술한 다원주의적 과학 공동체에 관한 논의, 그리고 동원의 책임에 대한 기존의 논의를 기반으로, 나는 인공지능 기반 과학 연구를 수행할 때 일부 과학자에게 다원주의적 공동체를 조성하기 위해 동료를 동원할 책임을 부여하자고 제안한다. 개인이 제시할 수 있는 관점에

26) Politi (2025), p. 7.; Fleisher and Šešelja (2023), p. 4.
27) Politi (2025), p. 8.

는 한계가 있으므로, 공동체 내에 다양한 관점이 적극적으로 제시되고 고려되도록 하기 위해서는 여러 사람의 노력이 필요하다. 내가 말하는 동원의 책임이란 이처럼 다양한 관점을 제시하고 고려하도록 동료를 설득할 책임이다.

나의 제안을 자세히 설명하기 위해 몇 가지를 명확하게 하겠다. 첫째, 나의 제안은 개인 과학자의 책임에 대한 것이지, 과학 공동체 자체의 책임에 대한 것은 아니다. 개인이 아닌 집단이 책임의 귀속 대상이 되는지에 대한 형이상학적 질문은 이 논문의 범위를 벗어난다. 둘째, 동원의 책임은 예견의 책임과 다르다. 모든 개별 과학자는 의도치 않은 결과에 대해 합당한 수준에서 예견할 책임이 있다. 반면, 동원의 책임은 과학 공동체 내에서 지도적 위치에 있거나 윤리 관련 업무를 담당하는 과학자가 예견의 책임과 함께 추가로 부담해야 하는 책임이다. 그리고 아마도 가장 중요한 세 번째 지점은, 동료를 동원한다는 것은 동료 개인의 연구 자체가 다원주의를 옹호하는 방향으로 진행되도록 설득해야 한다는 의미가 아니라는 점이다. 그보다는 공동체 전체에서 다양한 관점이 고려될 수 있도록 구조적인 노력이 필요하다는 뜻이다.

예를 들어 보겠다. 의료 진단을 위한 인공지능 모델을 개발한다고 해보자. 이때 일반적으로 제기될 수 있는 우려는 인공지능 모델이 환자의 질병을 오진할 때 발생할 여파에 관한 것이다. 그런데 인종차별의 문제에 관심이 많은 과학자라면 해당 인공지능 모델이 특정 인종에만 높은 오진율을 보일 수 있다고 우려할 수 있다.[28] 다원주의적 과학 공동체에서는 의료 진단 인공지능 모델이 불러올 여파에 대해 예견할 때, 이러한 인종차별에 대한 우려 역시 적극적으로 고려될 수 있어야 한다. 이를 위하여 공동체의 지도자 혹은 윤리 문제 담당자는 연구비 등을 할당하여 과학 연구에서 인종차별에 대한 의견이 적극적으로 개진되도록 환경을 조성할 수 있다. 예컨대 의료 진단 인공지능 모델의 인종 편향에 대한 분석에 자금을 지원하는 식으로 말이다. 이처럼 연구비 조성 등의 방법을 통하여 다양한 관점 제시에 기여할 수 있는 연구를 수행

28) 실제로 지금까지 서구 사회에서 발달한 의료 지식과 기술이 백인 남성의 몸에 맞춰져 있다는 문제의식이 제기된 바 있다. 관련 연구는 Fernández Pinto (2018) 참조.

하도록 동료 과학자를 독려 및 설득하는 것을, 동원의 책임의 예시로 볼 수 있다. 즉, 모든 과학자가 인종차별에 대해 연구할 필요는 없지만, 관련 연구와 관점이 적극적으로 개진될 수 있도록 해야 한다는 것이다. 이러한 노력을 통하여 3절에서 논의한 인공지능 기반 연구에서의 편향된 공동체 문제를 완화할 수 있을 것이다.

이와 같은 다원주의적 공동체는 4.1절에서 제안한 연구 프로젝트 중심의 맥락 의존적 기준을 검토하는 데도 효과적이다. 만약 프로젝트 외부의 인물이라면 해당 프로젝트 내에서는 고려되지 않았던 다른 관점을 바탕으로 해당 프로젝트에서 활용되는 합당한 예견의 기준을 비판적으로 검토할 수 있을 것이다. 프로젝트 내부의 인물이라면 이러한 비판적 검토가 상대적으로 어렵겠지만, 다원주의적 공동체에서의 다양한 관점을 인지한 상태에서 기준을 정립할 것이므로 기준의 편향성에 대한 우려가 경감될 것이다.

5. 나가며

이 논문에서 나는 더글러스의 주장을 비판적으로 검토하며 인공지능 시대 과학자의 도덕적 책임에 대해 논했다. 나는 과학자가 본인의 연구가 초래할 수 있는 의도치 않은 결과를 합당한 수준에서 예견할 책임이 있다는 더글러스의 주장에는 대체로 동의하면서도, 합당한 예견의 기준에 대해서는 더글러스와 다른 견해를 피력했다. 구체적으로는 첫째, 더글러스는 과학 공동체가 “합당한 예견”에 대한 단일한 기준을 제시할 것이라고 가정하지만, 나는 인공지능 기반 연구가 다학제적 특성을 보이므로 여러 기준이 존재할 수 있음을 강조했다. 둘째, 더글러스는 과학 공동체가 신속하게 기준을 설정할 수 있다고 보지만, 인공지능 기반 연구에서는 공동체가 합의에 이르는 속도보다 인공지능의 개발 및 활용 속도가 더 빠를 수 있음을 지적했다. 셋째, 공동체가 도덕적 판단의 기준이 된다는 더글러스의 견해와 달리, 나는 공동체 자체가 편향될 가능성이 있음을 강조했다. 그리고 이러한 세 가지 문제를 각각

“다학제의 문제”, “느린 기준 설정의 문제”, “편향된 공동체의 문제”라고 명명했다.

더글러스의 주장에 대한 이러한 비판을 바탕으로, 나는 다음 두 가지 사안을 제시했다. 첫째, 어디까지 예견하는 것이 합당한지 판단할 때는 공동체가 제시하는 단일하고 추상적인 기준에 전적으로 기대는 대신 연구 프로젝트가 수행되는 구체적인 맥락을 중심에 두자고 제안했다. 둘째, 공동체 내에서는 단일한 기준을 모색하는 대신 여러 관점이 적극적으로 고려되어야 하며, 이러한 다원주의적인 공동체를 조성하기 위하여 일부 과학자에게는 동료를 동원할 책임을 부여할 것을 제안했다.

이 논문은 과학철학에서의 과학과 가치 논의와 인공지능 윤리의 효과적인 접점을 모색했다는 의의가 있다. 과학과 비인식적 가치에 대한 논의는 21세기 과학철학의 거대한 한 축을 담당하고 있으며, 최근 국내에서도 관련 서적이 번역되는 등(엘리엇 2022) 이에 관한 관심이 증가하고 있다. 이 논문에서 보여주었듯, 과학과 가치에 대한 과학철학에서의 통찰을 적극적으로 활용한다면 인공지능의 실제 개발과 사용에 대해 보다 효과적이고 심도 있는 논의를 전개할 수 있을 것이다. 또한 반대로, 오늘날의 인공지능 기반 연구를 면밀하게 관찰한다면 과학과 가치에 대한 기존 논의를 현 상황에 걸맞게 더욱 발전시킬 수 있을 것이다.

참고문헌

고인석 (2012),「로봇이 책임과 권한의 주체일 수 있는가」,『철학논총』67권 1호, pp. 3-21.

목광수 (2010),「윤리적인 동물 실험의 철학적 옹호 가능성 검토」,『철학연구』90권, pp. 33-61.

엘리엇, 케빈 (2022),『과학에서 가치란 무엇인가: 연구 주제 선정부터 설계, 실행, 평가까지』, 김희봉 옮김, 김영사.

우태민, 신유정, 박범순 (2025),「인공지능 아실로마 회의와 기술결정론의 축적」,『과학기술학연구』24권 3호, pp. 4-42.

이상욱 (2011),「이해충돌과 과학 연구 윤리」,『과학철학』14권 11호, pp. 135-160.

이상형 (2016),「윤리적 인공지능은 가능한가? - 인공지능의 도덕적, 법적 책임 문제 -」,『법과정책연구』16권 4호, pp. 283-303.

최훈 (2009),「동물 신경 윤리: 동물 고통의 윤리적 의미」,『생명윤리』10권 2호, pp. 49-61.

최훈, 신중섭 (2007),「연구 부정행위와 연구 규범」,『과학철학』10권 2호, pp. 103-126.

한혜정 (2021),「파는 윤리: AI 윤리와 기업의 영향에 대하여」,『과학뒤켠』11호.

Aguirre, A. (2024), *The Pause Letter: One year later*, Future of Life Institute, https://futureoflife.org/ai/the-pause-letter-one-year-later/ (accessed 12 Feb 2025).

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016), “Machine Bias. There’s Software Used across the Country to Predict Future Criminals. And It’s Biased against Blacks”, *ProPublica*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 12 Feb 2025).

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2022), “Machine Bias”,

In Martin, K. (ed.), *Ethics of Data and Analytics: Concepts and Cases*, Auerbach Publications, pp. 254-64.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., ... and Mindermann, S. (2024), "Managing extreme AI risks amid rapid progress". *Science* 384: pp. 842-5.

Bridgman, P. W. (1947), "Scientists and Social Responsibility", *The Scientific Monthly* 65: pp. 148-54.

Bush, V. (1945), *Science: The endless frontier: A Report to the President on a Program for Postwar Scientific Research*, National Science Foundation.

Clark, E., and Khosrowi, D. (2022), "Decentring the Discoverer: How AI Helps us Rethink Scientific Discovery", *Synthese* 200: p. 463.

Criddle, C. (2024), "AI Groups Rush to Redesign Model Testing and Create New benchmarks", *Financial Times*, https://www.ft.com/content/866ad6e9-f8fe-451f-9b00-cb9f638c7c59 (accessed 12 Feb 2025).

De Langhe, R. (2009), "Why Should I Adopt Pluralism", in Garnett, R., Olsen, E., and Starr, M. (eds.), Economic *Pluralism*, Routledge, pp. 109-20.

Douglas, H. E.(2000), "Inductive Risk and Values in Science", *Philosophy of science* 67: pp. 559-79.

_____ (2003), "The Moral Responsibilities of Scientists (Tensions between Autonomy and Responsibility)", *American Philosophical Quarterly* 40: pp. 59-68.

_____ (2009), *Science, policy, and the value-free ideal*, University of Pittsburgh Press.

_____ (2021), "Scientific Freedom and Social Responsibility", In Hartl, H. and Tuboly, A. T. (eds.), *Science, freedom, democracy*, Routledge, pp. 68-87.

Duede, E. (2023), "Deep Learning Opacity in Scientific Discovery", *Philosophy of Science* 90: pp. 1089-99.

Elliott, K. C., and McKaughan, D. J. (2009), “How Values in Scientific Discovery and Pursuit Alter Theory Appraisal”, *Philosophy of Science* 76: pp. 598-611.

European Medicines Agency [EMA]. (2024), *Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle*, EMA/CHMP/CVMP/83833/2023.

Fazelpour, S., and Danks, D. (2021), “Algorithmic bias: Senses, sources, solutions”, *Philosophy Compass* 16.

Feinberg, J. (1970), *Doing and deserving*, Princeton University Press.

Fernández Pinto, M. (2018), “Democratizing Strategies for Industry-funded Medical Research: A Cautionary Tale”, *Philosophy of Science* 85: pp. 882-94.

Fleisher, W., and Šešelja, D. (2023), “Responsibility for Collective Epistemic Harms”, *Philosophy of Science* 90: pp. 1-20.

Food and Drug Administration [FDA]. (2025), *Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products: Draft Guidance for Industry and Other Interested Parties*, FDA-2024-D-4689.

Future of Life Institute (2023), *Pause Giant AI Experiments: An Open Letter*, https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (accessed 12 Feb 2025)

Griffen, E. J., Dossetter, A. G., and Leach, A. G. (2020), “Chemists: AI is Here; Unite to Get the Benefits”, *Journal of Medicinal Chemistry* 63: pp. 8695-704.

Hagendorff, T. (2020), “The Ethics of AI Ethics: An Evaluation of Guidelines”, *Minds & Machines* 30: pp. 99-120.

Hicks, M. T., Humphries, J., and Slater, J. (2024), “ChatGPT is Bullshit”, *Ethics and Information Technology* 26.

Hindriks, F. (2019), “The Duty to Join Forces: When Individuals Lack

Control", *The Monist* 102: pp. 204-20.

Holman, B., and Elliott, K. C. (2018), "The Promise and Perils of Industry-funded science", *Philosophy Compass* 13.

Kitcher, P. (2001), *Science, Truth, and Democracy*, Oxford University Press.

_____ (2011), *Science in a Democratic Society*, Prometheus Books.

Kuhlman, C., Jackson, L., and Chunara, R. (2020), "No Computation Without Representation: Avoiding Data and Algorithm Biases Through Diversity", *arXiv preprint arXiv:2002.11836*.

Lefevere, M., and Schliesser, E. (2014), "Private Epistemic Virtue, Public Vices: Moral Responsibility in the Policy Sciences", in Martini, C., and Boumans, M. (eds.), *Experts and Consensus in Social Science*, Springer International Publishing, pp. 275-95.

Lloyd, E. A. (2005), *The Case of the Female Orgasm: Bias in the Science of Evolution*, Harvard University Press.

Longino, H. E. (1990), *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

_____ (2002), *The fate of knowledge*. Princeton University Press.

MacLeod, M. (2018), "What Makes Interdisciplinarity Difficult? Some Consequences of Domain Specificity in Interdisciplinary Practice", *Synthese* 195: pp. 697-720.

MacLeod, M., and Nersessian, N. J. (2013), "Coupling Simulation and Experiment: The Bimodal Strategy in Integrative Systems Biology", *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44: pp. 572-84.

_____ (2014), "Strategies for Coordinating Experimentation and Modeling in Integrative Systems Biology", *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 322: pp. 230-39.

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., ... and Clark, J., (2024), *The AI Index 2024 Annual Report*, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

Mock, M., Edavettal, S., Langmead, C., and Russell, A. (2023), "AI can Help to Speed Up Drug Discovery—But Only If We Give It the Right Data", *Nature* 621: pp. 467-70.

Mullin, R. (2023), "The Tricky Ethics of AI in the Lab", *Chemical & Engineering News* 101: pp. 30-5.

National Academy of Science [NAS] (1992), *Responsible Science: Ensuring the Integrity of the Research Process*, National Academies Press.

Nature (2023), "AI's Potential to Accelerate Drug Discovery Needs a Reality Check", *Nature* 622.

Ochigame, R. (2022), "The Invention of 'Ethical AI': How Big Tech Manipulates Academia to Avoid Regulation", in Phan, T., Goldenfein, J., and Kuch, D. (eds.), *Economies of Virtue − The Circulation of 'Ethics' in AI*, Institute of Network Cultures.

Politi, V. (2025), "The Collective Responsibilities of Science: Toward a Normative Framework", *Philosophy of Science* 92(1): pp. 1-18.

Rudner, R. (1953), "The Scientist qua Scientist Makes Value Judgments", *Philosophy of Science* 20: pp. 1-6.

Scalet, S. P. (2003), "Fitting the People They are Meant to Serve: Reasonable Persons in the American Legal System", *Law and Philosophy* 22: pp. 75-110.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. (2024), "AI Models Collapse When Trained on Recursively Generated Data", *Nature* 631: pp. 755-59.

Simkute, A., Luger, E., Evans, M., and Jones, R. (2024), "It Is There, and You Need It, So Why Do You Not Use It? Achieving Better

Adoption of AI Systems by Domain Experts, in the Case Study of Natural Science Research", *arXiv preprint arXiv:2403.16895.*

Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. (2022), "Dual Use of Artificial-intelligence-powered Drug Discovery", *Nature machine intelligence* 4: pp. 189-91.

Young, E., Wajcman, J., and Sprejer, L. (2023), "Mind the Gender Gap: Inequalities in the Emergent Professions of Artificial Intelligence (AI) and Data Science", New *Technology, Work and Employment* 38: pp. 391-414.

논문 투고일	2025. 02. 12.
심사 완료일	2025. 03. 11.
게재 확정일	2025. 03. 23.

The Moral Responsibilities of Scientists in the Age of Artificial Intelligence: On the Standard of Reasonable Foresight

HyeJeong Han

About two decades ago, Heather Douglas argued in her article “The Moral Responsibilities of Scientists” that scientists have a moral responsibility to foresee the unintended consequences of their work and that the standard for reasonable foresight should be determined by the scientific community. The aim of this paper is to critically examine Douglas’s argument and thereby to explore the moral responsibilities of scientists in the age of artificial intelligence. I argue that Douglas’s framework fails to adequately capture several key aspects of AI-based scientific research, particularly what I call the “multidisciplinarity problem”, the “slow standard-setting problem”, and the “biased community problem”. To address these issues, I propose two solutions: first, the standard of reasonable foresight should be settled centring on the specific contexts of research projects; second, at least some scientists should take on the responsibility of mobilizing their colleagues to promote a pluralistic community that actively considers diverse perspectives.

Keywords: reasonable foresight, unintended consequences, scientific community, community standard, context-dependence, pluralism, Heather Douglas