

인공지능으로 인한 과학적 불확실성의 소통 및 관리를 위한 전략적 고찰

성두현(rcdsung@gmail.com)

한국과학기술원(KAIST) 디지털인문사회과학센터

연세대학교 언더우드국제대학

연사 소개

- **학력:**

- KAIST 물리학과 학사.
- KAIST 과학기술정책대학원 석사.
- 케임브리지 대학교 과학사 및 과학철학과 방문학생.
- KAIST 과학기술정책대학원 박사.

- **경력:**

- (현) 연세대학교 언더우드국제대학 강사.
- 라이프니츠 하노버 대학교 SOCRATES 연구소 박사후 연구원.
- (현) KAIST 디지털인문사회과학센터 박사후 연구원.

연사 소개

- **연구 분야:** 일반과학철학(과학과 가치), 사회인식론, 의학철학.
 - 과학적 추론 및 의사결정의 합리성 및 윤리성(Sung & Holman, 2023).
 - 과학적 근거의 다원성(Holman & Sung, 2025; Sung & Holman, 2023).
 - 객관적이고 윤리적인 과학을 위한 사회제도 설계(Sung, 2026; Sung & Holman, 2023).
- **주요 출판물:**
 - **Sung, D., & B. Holman. (2023).** Against evidential pluralism in pharmaceutical regulation. *Philosophy of Science*, 90(5), 1276–1285.
 - **Holman, B., & D. Sung. (2025).** Evidence in medicine. In A. Broadbent (Ed.) *The Oxford handbook of philosophy of medicine* (pp. 448–470). Oxford University Press.
 - **Sung, D. (2026).** On epistemic risks in research method validation. *Studies in History and Philosophy of Science*, 119.

강연 주제

- **주제:** 인공지능이 과학탐구 과정에서 발생하는 인식적 우발사태의 소통 및 관리에 미치는 (잠재적) 파급효과.
 - (잠재적) 파급효과에 대한 대처를 위한 **메타적/전략적** 고찰.

강연 주제

- **참조 논문:** Sung, D. (2026a). 항공우주의학 연구개발 및 규제 의사결정을 위한 철학적 길잡이. *The Korean Journal of Aerospace and Environmental Medicine*, 36(2), 34–40.
- **유용한 배경지식:**
 - 의사결정 이론(decision theory).
 - 위험관리 이론(risk management theory).
 - 언어행위 이론(speech act theory).
 - 제한된 합리성(bounded rationality).
 - 임상시험 연구(clinical trial research).

강연 순서

- 강연은 **총 3부**로 구성.
- **핵심어**(keyword).
 - 특강 1부 및 2부는 핵심어(영문 번역 포함)를 소개함.
- **순서:**
 1. 과학적 불확실성의 소통 및 관리.
 2. 인공지능의 잠재적 파급효과.
 3. 연구 전략 및 자원.

1부: 과학적 불확실성의 소통 및 관리

1.1 개요

- 주제: 과학적 불확실성의 소통 및 관리.
 - 합리적 의사결정의 문제로서 논의.
- 목적: 인공지능이 영향을 미칠 수 있는 과학탐구의 핵심 활동 중 하나(과학적 가설 평가 및 주장)을 조명.
- 핵심어:
 - 신념의 윤리학(ethics of belief).
 - 인식적 우발사태(epistemic contingency).
 - 귀납적 위험(inductive risk).
 - 과학적 주장(scientific assertion).

1.1 개요

- **문제:** 과학적 탐구가 필연적으로 수반하는 불확실함을 어떻게 소통하고 관리해야 하는가?
 - **관점 1(인지):** 의사결정 이론.
 - **관점 2(소통):** 언어행위 이론.

1.2 의사결정 이론이란?

- **합리적인 의사결정의 일반적인 조건을 탐구하는 학제간 분야.**
 - (응용)수학, 철학, 경제학, 사회과학, 심리학 등을 포함함.
 - **설명의 역할과 규범의 역할**을 검할 수 있음.
- 의사결정 이론의 기본 요소:
 - **선택지.**
 - 선택지의 결과에 대응하는 **확률.**
 - 선택지의 결과에 대응하는 **효용.**
 - 합리적인 선택을 위한 원칙 혹은 규범
 - 예: 기대효용의 극대화(expected utility maximization).

1.2 의사결정 이론이란?

- 탐구 방식: 역사적 예시 및 논증 분석.
 - 역사적 예 1: 파스칼의 내기.
 - 역사적 예 2: 신념의 윤리학과 실행주의(pragmatism).
 - 논증 분석: 귀납적 위험 논증과 과학적 객관성.

1.3. 파스칼의 내기

- 프랑스 철학자 **블레즈 파스칼**(Blaise Pascal)이 <**팡세**(Pensée)>에서 소개한 신의 존재에 대한 논증.
- **형이상학적 토대**에서 출발하여 연역적 추론을 통해 신의 존재를 **증명**하는 존재적 논증과 달리 **신앙의 합리성**을 수학적으로 **조명**한다는 점에서 독보적임.
- (현대) 의사결정이론의 첫 예로서 알려져 있음.

1.3. 파스칼의 내기

- 선택지:
 - H: 유신론.
 - $\sim H$: 무신론.
- 확률 P:
 - $P(H)$: 신이 존재할 확률.
 - $P(\sim H)$: 신이 부재할 확률.
 - $P(H) + P(\sim H) = 1$.
- 효용 U:
 - $U(H)$: 신이 존재하는 실존적 상황의 효용.
 - $U(\sim H)$: 신이 부재하는 실존적 상황의 효용.
- 기대효용 E: 모든 선택지에 대응하는 모든 확률 및 효용의 곱의 총합.

1.3. 파스칼의 내기

- 신의 존재 여부에 따른 유신론과 무신론의 효용.

	H는 참/ $P(H) = 1$	H는 거짓/ $P(H) = 0$
유신론을 수용함	$U_H(H) = +\infty$ (구원을 받음)	$U_H(\sim H)$
무신론을 수용함	$U_{\sim H}(H) = -\infty$ (구원을 포기함)	$U_{\sim H}(\sim H)$

1.3. 파스칼의 내기

- 유신론과 무신론의 기대효용:

- $E_H = P(H) \times U_H(H) + P(\sim H) \times U_H(\sim H).$
- $E_{\sim H} = P(H) \times U_{\sim H}(H) + P(\sim H) \times U_{\sim H}(\sim H).$

- $P(H) = 1$ 인 경우:

- $E_H = 1 \times U_H(H) + 0 \times U_H(\sim H) = U_H(H) = +\infty.$
- $E_{\sim H} = 1 \times U_{\sim H}(H) + 0 \times U_{\sim H}(\sim H) = U_{\sim H}(H) = -\infty.$

- $P(H) = 0$ 인 경우:

- $E_H = 0 \times U_H(H) + 1 \times U_H(\sim H) = U_H(\sim H).$
- $E_{\sim H} = 0 \times U_{\sim H}(H) + 1 \times U_{\sim H}(\sim H) = U_{\sim H}(\sim H).$

1.3. 파스칼의 내기

- 두 기대효용 범위의 차이:
 - 유신론: $U_H(\sim H) \leq E_H \leq +\infty$.
 - 무신론: $-\infty \leq E_{\sim H} \leq U_{\sim H}(\sim H)$.
- 결론: 유신론을 수용하는 것이 합리적임.

1.3. 파스칼의 내기

- 교훈:

- (인식적) 확신과 합리성은 별개의 문제임.
- 확실한 근거가 없어도 합리적인 의사결정은 가능함.

1.4. 신념의 윤리학

- **문제:** 확실한 근거가 없는 상황에서 신념을 과연 정당할 수 있을까?

1.4. 신념의 윤리학

- **윌리엄 클리포드**(William K. Clifford)가 1877년에 <신념의 윤리학(The Ethics of Belief)>에서 주장한 **근거우선주의**(strong evidentialism).
- **근거우선주의**: 확실한 근거가 부재한 상황에서 신념을 가지는 것은 (인식적으로나 도덕적으로나) 정당하지 않음.
 - (지성적) 범죄와 같은 본질의 부정당함.

1.4. 신념의 윤리학

- 윌리엄 제임스(William James)가 1896년에 <신념의 의지(The Will to Believe)>에 주장한 **실행주의**(pragmatism).
- 다음 조건하에 신념을 가지는 것은 정당하거나 의무적임.
 - 인식적 우발성(epistemic contingency). 인식적 확신은 불가능함.
 - 책무성(commitment). (도덕적) 책무에 준하는 결정을 내려야 함.
 - 결단성(decisiveness). 주어진 대안 중 오로지 하나를 선택해야 함.
 - 중대성(significance). 모든 대안이 (도덕적으로) 중대함.
- 위 조건이 충족됨에도 불구하고 인식적 확신을 강요하는 것은 오히려 부당할 수 있음.

1.4. 신념의 윤리학

- 예시: 파스칼의 내기와 종교적 신념의 윤리성.
- 다음 조건하에 신념을 가지는 것은 정당하거나 의무적임.
 - 인식적 우발성. 인간의 경험과 지성만으로 신의 존재를 확증할 수 없음.
 - 책무성. 유신론 혹은 무신론은 책무(신앙 및 헌신)의 문제임.
 - 결단성. 유신론과 무신론은 양립할 수 없음.
 - 중대성. 유신론 혹은 무신론의 수용은 도덕적으로 중대한 결과를 낳음.
- 위 조건이 충족됨에도 불구하고 종교적 신념을 부도덕하게 해석할 수 있는 근거우선주의는 종교의 자유와 상충됨.

1.4. 신념의 윤리학

- **교훈:**

- 결정적인 근거 없음에도 정당한 신념을 가질 수 있는 경우는 **적어도 하나가 존재함**(윌리엄 제임스의 **실행주의** 논증).
- (인식적) 확신에 대한 강조가 합리적인 의사결정 및 실천과 대립할 가능성이 존재함.

1.5. 귀납적 위험

- **문제:** 과학적 가설 평가.
 - 과학적 추론 과정에서 **가치판단**을 내리는 것은 정당한가?
- 과학철학의 두 고전.
 - **리처드 러드너**(Richard Rudner)의 <과학자는 과학자로서 가치판단을 내린다(The Scientist Qua Scientist Makes Value Judgments)>(1953).
 - 윌리엄 제임스의 논증과 역사적 관련성이 있음(Magnus, 2022; Rescher, 1977).
 - **헤더 더글라스**(Heather Douglas)의 <귀납적 위험과 가치판단이 과학 탐구에서 맡는 역할(Inductive Risk and Values in Science)>(2000).

1.5. 귀납적 위험

- **문제:** 과학적 불확실성이 존재할 때 가치판단을 내리는 것은 정당한가?
- **귀납적 오류의 위험.**
 - 귀납적 위험은 인식적 우발성의 한 경우임.
 - **제1종오류(false positive):** 거짓된 가설을 참으로서 수용하는 오류.
 - **제2종오류(false negative):** 참된 가설을 거짓으로서 수용하는 오류.
 - **통계학적으로 제1종오류의 위험과 제2종오류의 위험 사이의 상충관계(tradeoff)**가 존재함. 둘을 동시에 최소화하는 것은 불가능함.
 - 시간 및 자원의 한계없이 근거를 확보하는 것 또한 현실적으로 불가능함.

1.5. 귀납적 위험

- 가설 H: '약물 P는 안전하다.'

	가설 H는 참	가설 H는 거짓
가설 H를 수용	올바른 가설 수용	제1종오류
가설 H를 거부	제2종오류	올바른 가설 거부

	가설 H는 참	가설 H는 거짓
가설 H를 수용	올바른 가설 수용	제1종오류로 인한 위험 (예: 보건위기, 규제에 대한 불신 등)
가설 H를 거부	제2종오류로 인한 위험 (예: 치료 기회 박탈, 자원 낭비, 규제에 대한 불신 등)	올바른 가설 거부

1.5. 귀납적 위험

- 리처드 러드너의 귀납적 위험 논증.
 - 인식적 우발성(epistemic contingency). 귀납적 위험이 존재함.
 - 책무성(commitment). 가설 수용은 인식적 책무의 문제임.
 - 결단성(decisiveness). 가설은 동시에 수용하고 거부할 수 없음.
 - 중대성(significance). 가설 수용 및 거부는 (도덕적으로) 중대한 결과를 낳음.
- 과학자는 두 귀납적 오류로 인한 결과의 (상대적) 중대성에 대한 가치판단을 내릴 수 밖에 없음.
- 결론: “과학자는 과학자로서 가치판단을 내린다.”

1.5. 귀납적 위험

- 헤더 더글라스(2000, 2009)의 귀납적 위험 논증.
 - 귀납적 위험은 과학 추론 뿐만 아니라 과학 추론에 **(직간접적인) 영향**을 미치는 활동(예: 표본 분석, 통계모형 선택 등)을 할 때도 존재함.
 - 과학자는 과학자로서 상당한 지성적, 사회적, 정치적 **권위**를 누림.
 - 과학자는 과학자라는 이유로 보편적인 도덕적 의무에서 **면책권**을 누릴 수 없음.
- **결론:** “과학자는 과학자로서 가설을 평가할 때 **예상이 가능한** 귀납적 위험에 대한 도덕적 판단을 내릴 **의무**가 있다.”
 - 다만 **가치판단**은 근거와 본질적으로 다르며, **근거의 충분함**을 판단하는 기준의 역할만 맡을 수 있음.

1.5. 귀납적 위험

- 가설 H: '약물 P는 안전하다.'

	가설 H는 참	가설 H는 거짓
가설 H를 수용	올바른 가설 수용	제1종오류
가설 H를 거부	제2종오류	올바른 가설 거부

	가설 H는 참	가설 H는 거짓
가설 H를 수용	올바른 가설 수용	제1종오류로 인한 위험 (예: 보건위기, 규제에 대한 불신 등)
가설 H를 거부	제2종오류로 인한 위험 (예: 치료 기회 박탈, 자원 낭비, 규제에 대한 불신 등)	올바른 가설 거부

1.6. 귀납적 위험의 소통

- 귀납적 위험은 **인지**(cognition)의 문제일 뿐만 아니라 **소통**(communication)의 문제임(John, 2015).
- 언어행위로서의 **과학적 주장**.
 - 과학적 주장은 가설의 전달을 넘어 청중에게 (의도되거나 의도치 않은) 영향을 미침.
 - 과학자는 단순히 가설에 대해 **추론**하는 것을 넘어 다수의 청중을 대상으로 과학적 가설을 **주장**함.

1.6. 귀납적 위험의 소통

- **소통윤리학**(ethics of communication)의 원칙 및 개념적 구분 (John, 2015).
 - **임마누엘 칸트**(Immanuel Kant)의 **의무능력원칙**(ought-implies-can principle): 도덕적 주체가 실천할 수 있는 규범만이 효력을 가짐.
 - 소통의 두 맥락:
 - 청중의 특성을 완전히 예상할 수 있는 **사적소통**(private communication).
 - 청중의 특성을 완전히 예상할 수 없는 **공적소통**(public communication).

1.6. 귀납적 위험의 소통

- 과학자들은 **사적소통**(예: 학술대회 발표)의 경우에는 주장하는 행위가 청중에게 미칠 도덕적 여파를 충분히 예상할 수 있으나, **공적소통**(예: 학술지 논문 게재)의 경우에는 도덕적 여파를 충분히 예상할 수 없음.
 - 헤더 더글라스의 주장을 **공적소통**에 적용하는 것은 **의무능력원칙**을 위반할 소지가 있음.
 - 모든 과학자들이 개별적으로 도덕적 판단을 내리는 것이 (정치적으로) 바람직한가라는 문제도 있음.
- **대안(John, 2015)**: 모든 과학자가 **개별적으로** 가치판단을 내리는 대신 **제도설계**를 통해 **공통된** 인식적 기준을 확립할 필요가 있음.

1.7. 정리

- **교훈:**

- 과학적 불확실성의 소통 및 관리는 **특수한 유형**의 의사결정 문제임.
- 과학적 불확실성에도 불구하고 합리적이거나 도덕적인 의사결정 및 실천은 **가능함**.
- 과학적 불확실성의 소통 및 관리를 위한 **제도설계**는 중대한 문제임.

2부: 인공지능의 잠재적 파급효과

2.1. 개요

- **주제:** 인공지능으로 인한 (잠재적) 파급의 층위.
 - 층위 1: (귀납적 위험을 포함한) 인식적 위험 일반의 분류.
 - 층위 2: 제한된 합리성과 인공지능의 (존재적) 지위.
- **목적:** 인공지능의 파급효과에 대한 **선제연구(preemptive research)** 소개.
 - Sung (2026a, 2026b, under review).
 - Sung & Holman (2023).
- **핵심어:**
 - 인식적 위험(epistemic risk).
 - 제한된 합리성(bounded rationality).

2.1. 개요

- 2부 논의의 전제조건:
 - ‘인공지능’과 ‘대형언어모형(large language model)’은 동의어임.
 - 인공지능의 역할은 과학적 추론 혹은 소통을 부분적으로 혹은 일체 **자동화**하는 것임.
 - 인공지능은 본질적으로 블랙박스(black box)임. 즉, 인공지능의 인지적 혹은 소통적 수행능력에 대한 **인과적 설명**은 불가능함.
 - 인공지능을 **인간의 가치판단 혹은 편향**으로부터 자유롭게 설계하는 것은 본질적으로 불가능함(Biddle, 2026).

2.2. 인공지능과 인식적 위험의 분류

- 헤더 더글라스(2000, 2009)의 귀납적 위험 논증.
 - 귀납적 위험은 과학 추론 뿐만 아니라 과학 추론에 (직간접적인) 영향을 미치는 활동(예: 표본 분석, 통계모형 선택)을 할 때도 존재함.
 - 과학자는 상당한 지성적, 사회적, 정치적 권위를 누림.
 - 과학자는 과학자라는 이유로 보편적인 도덕적 의무에서 면책권을 누릴 수 없음.
- 귀납적 위험과 인식적 위험 사이의 개념적 관계(Sung, under review).
 - 인식적 위험의 분류, 소통과 관리.
 - 위험관리에 있어서 인과적 책임 및 도덕적 책임 사이의 구분.
 - 사례탐구: 임상시험 과정에서 발생하는 인식적 위험의 소통 및 관리.

2.3. 인공지능과 제한된 합리성

- 제한된 합리성과 의사결정이론의 **인식론적 전제**에 대한 비판.
 - 인간과 인공지능은 과연 의사결정 이론이 상정하는 **전지적 합리성**에 도달할 수 있는가?(Wheeler, 2026).
 - 전지적 합리성은 애초에 물리적으로 가능한가?
- 의사결정 이론의 기본 요소 및 비판의 층위:
 - 의사결정 주체의 특성 및 환경적 요인.
 - **층위 1: 선택지.**
 - **층위 2: 선택지의 결과에 대응하는 확률.**
 - **층위 3: 선택지의 결과에 대응하는 효용.**
 - **층위 4: 합리적인 선택을 위한 원칙 혹은 규범(예: 기대효용의 극대화).**

2.4. 정리

- **교훈:**

- 과학적 불확실성의 소통 및 관리는 **특수한 유형**의 의사결정 문제임.
- 과학적 불확실성에도 불구하고 합리적이거나 도덕적인 의사결정 및 실천은 **가능함**.
- 과학적 불확실성의 소통 및 관리를 위한 **제도설계**는 중대한 문제임.

3부: 연구 전략 및 자원

3.1. 개요

- **목적:** 특강 주제에 대한 전략적 고찰을 위한 문헌.
 - **실행주의 철학, 과학적 불확실성의 소통 및 관리 혹은 인공지능의 존재적 지위에 관한 저서.**
 - 우리말로 저술되거나 옮겨진 저서.
- 상세한 주제는 **스탠포드 철학 백과사전(SEP)**의 항목을 바탕으로 고찰할 수 있음.
 - **검색어 예:** 'Decision theory,' 'Pascal's Wager,' 'Risk,' 'Expected utility,' 'Bounded Rationality,' 'Scientific objectivity,' 'Ethics of AI' 등.

3.2. 저서 소개

- <과학에서 가치란 무엇인가(A Tapestry of Values)>
 - 케빈 엘리엇(Kevin C. Elliott) 저, 김희봉 옮김.
 - 김영사.
- 탐독을 위한 질문:
 - 과학에서 가치란 무엇인가?
 - 과학 탐구 중에 내리는 가치판단은 어떠한 경우에 정당한가?
 - 과학은 사회와 어떠한 관계를 맺어야 하는가?

3.2. 저서 소개

- <새로운 실용주의 과학철학(Realism for Realistic People)>
 - 장하석 저, 전대호 옮김.
 - 김영사.
- 탐독을 위한 질문:
 - 철학자가 보는 과학의 현실은 어떠한가?
 - 철학은 어떻게 과학에 기여할 수 있을까?

3.2. 저서 소개

- <마음의 개념(The Concept of Mind)>
 - **Gilbert Ryle**(길버트 라일) 저, 이한우 옮김.
 - 문예출판사.
- **탐독을 위한 질문:**
 - 마음은 과연 무엇인가?
 - 범주오류(category mistake)란 무엇인가?
 - 범주오류의 원인은 무엇인가?

3.2. 저서 소개

- <허버트 사이먼과 사고의 재설계>
 - 배인철 저.
 - 도서출판 해남(2026년 10월 12일 1쇄 발행).
- **탐독을 위한 질문:**
 - 제한된 합리성의 개념은 어떠한 (학제간) 연구의 결실인가?
 - 제한된 합리성의 개념을 통해 인간의 지능과 인공지능을 어떻게 이해해야 하는가?
 - 지능과 사회 사이의 관계는 어떠한가? 지능과 사회 사이의 관계는 어떠해야 하는가?

3.3 맺으며

- **주제:** 인공지능이 과학탐구 과정에서 발생하는 인식적 우발사태의 소통 및 관리에 미치는 (잠재적) 파급효과.
 - (잠재적) 파급효과에 대한 대처를 위한 **메타적/전략적** 고찰.

Thank you

참고문헌

- **Biddle, J.B. (2026).** Artificial intelligence: Values, governance and policy. In K.C. Elliott, & T. Richards (Eds.) *The Routledge handbook of values and science* (pp. 511–522). Routledge.
<https://doi.org/10.4324/9781003469100-51>.
- **Douglas, H. (2000).** Inductive risk and values in science. *Philosophy of Science*, 67(4), 559–579.
<https://doi.org/10.1086/392855>.
- **Douglas, H.E. (2009).** *Science, policy, and the value-free ideal*. University of Pittsburgh Press.
- **John, S. (2015).** Inductive risk and the contexts of communication. *Synthese*, 192, 79–96.
<https://doi.org/10.1007/s11229-014-0554-7>.
- **Magnus, P.D. (2022).** William James on risk, efficacy, and evidentialism. *Episteme*, 19(1), 146–158. <https://doi.org/10.1017/epi.2020.17>.

참고문헌

- **Rescher, N. (1977).** *Methodological pragmatism: A systems-theoretic knowledge to the theory of knowledge*. Blackwell.
- **Rudner, R. (1953).** The scientist qua scientist makes value judgments. *Philosophy of Science*, 20(1), 1–6. <https://doi.org/10.1086/287231>.
- **Sung, D. (2026a).** Decision-making in aerospace medicine research, development, and regulation: A philosophical primer. *The Korean Journal of Aerospace and Environmental Medicine*, 36(2), 34–40. <https://doi.org/10.46246/KJAsEM.260010>.
- **Sung, D. (2026b).** On epistemic risks in research method validation. *Studies in History and Philosophy of Science*, 119. <https://doi.org/10.1016/j.shpsa.2026.102206>.
- **Sung, D. (Under review).** A pragmatist approach to conceptual engineering of inductive risk. *Synthese*.

참고문헌

- **Sung, D., & B. Holman. (2023).** Against evidential pluralism in pharmaceutical regulation. *Philosophy of Science*, 90(5), 1276–1285. <https://doi.org/10.1017/psa.2023.40>.
- **Wheeler, G. (2026).** All rationality is bounded. *Synthese*, 207. <https://doi.org/10.1007/s11229-026-05625-7>.