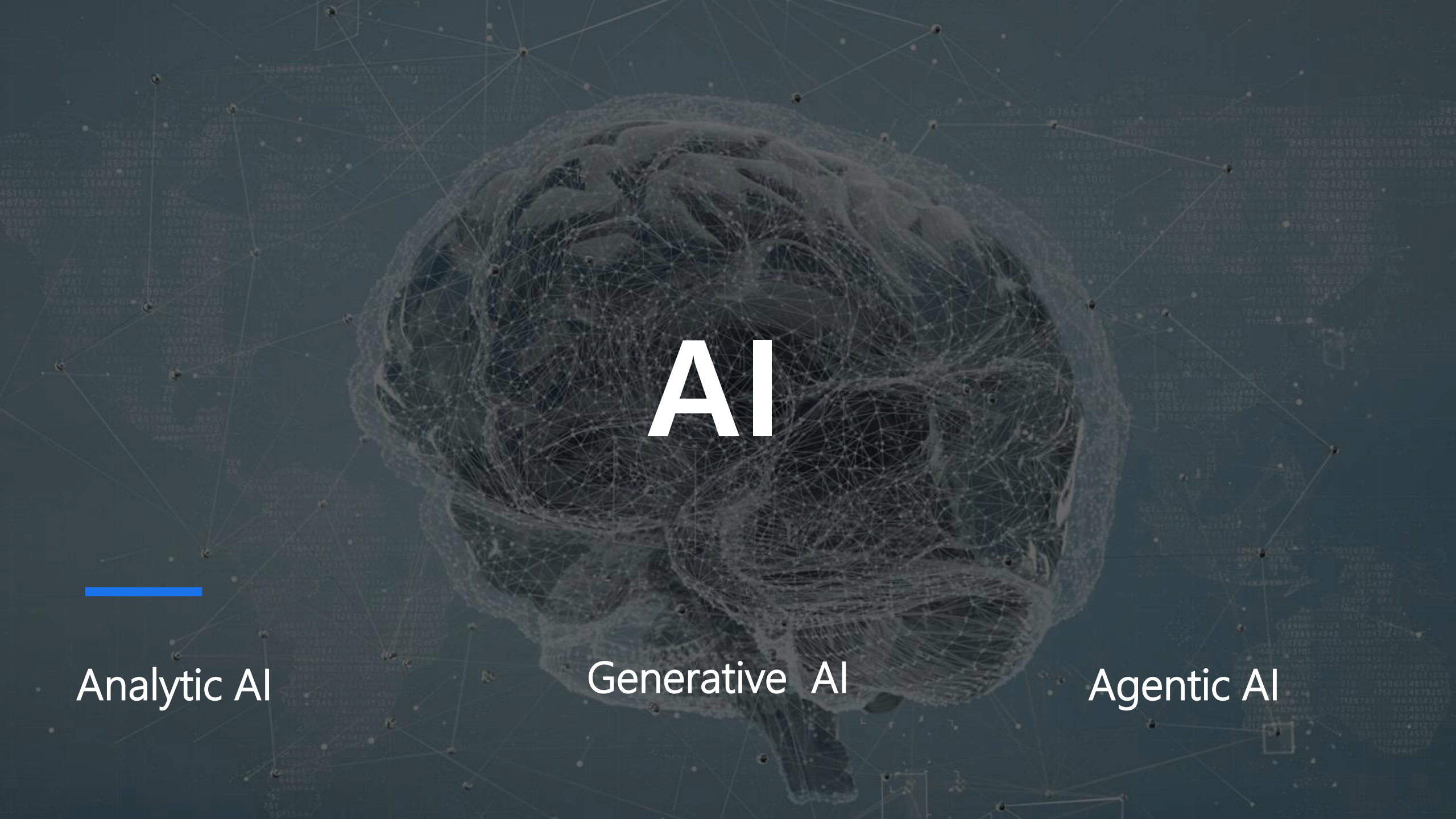
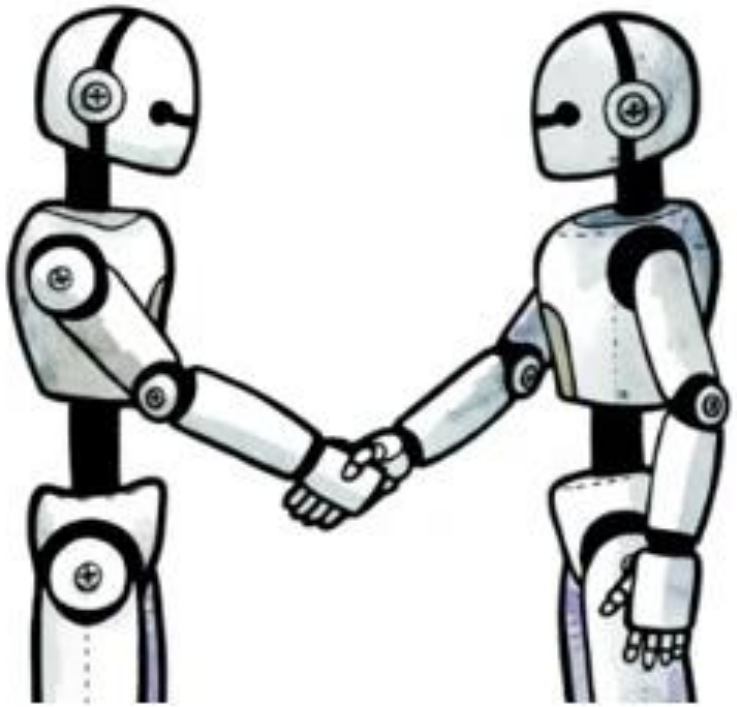
서울대학교 경영대학
박소정

AI와 보험산업: 신뢰, 공정성, 그리고 사람


AI
Analytic AI
Generative AI
Agentic AI

Imagine, Agentic AI



- Intake Agent
- Risk Profiling Agent
- Pricing & Product Agent
- Compliance & Fairness Agent
- Decision Orchestrator Agent
- Learning and Feedback Agent

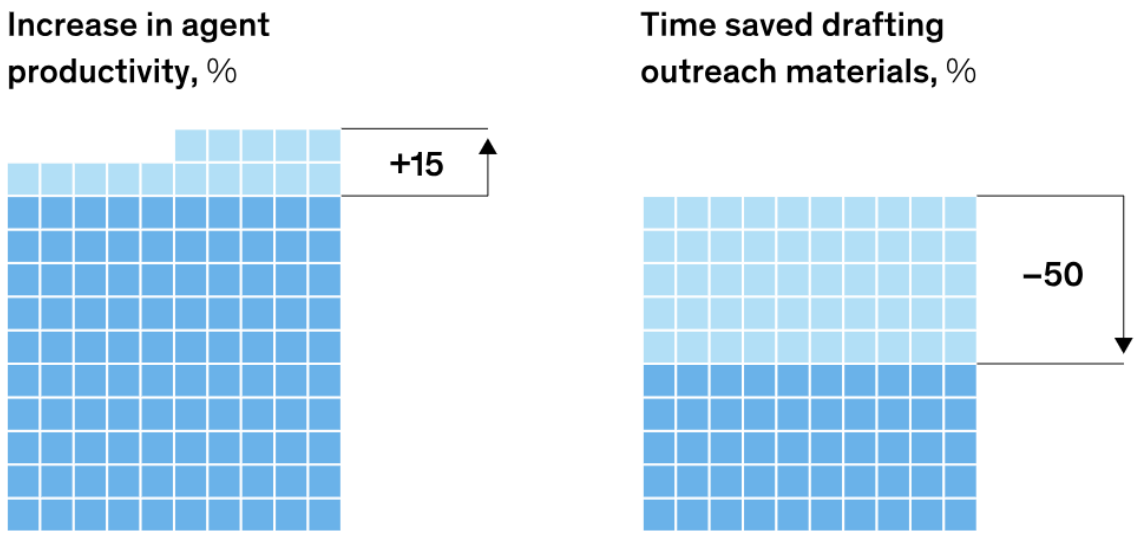
아래 목록 중 이질적인 것 하나는?

- Zara, Kate, Ada, Allie, Tomi, Dr Jeevan, Chatty, Botty, Jarvis or ABIe

- 챗봇 for Zurich, GEICO, Axa, Allianz, Tokyo Marine, MetLife, Generali, China Life and Allstate respectively – Botty is a Swiss shoe store chain

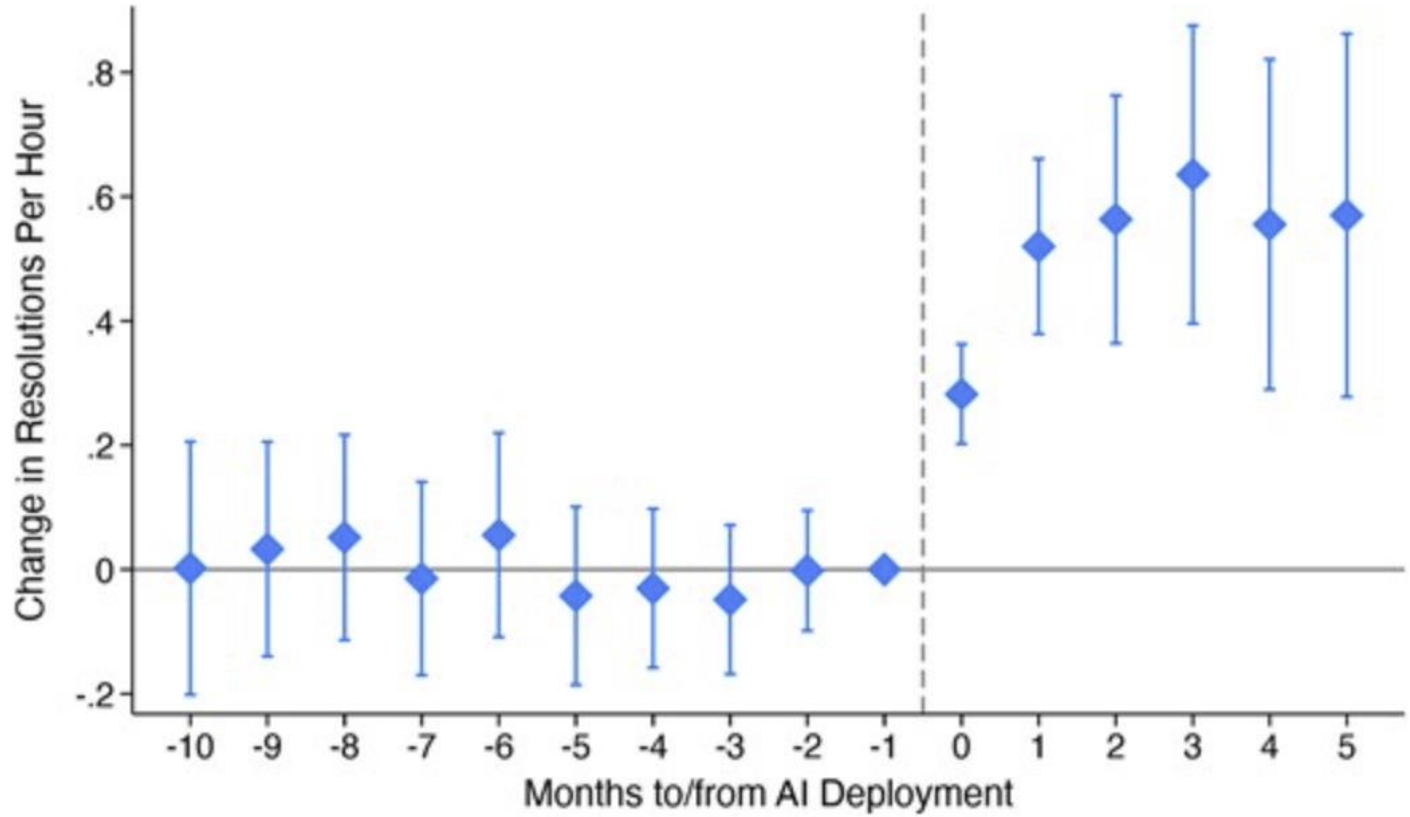
현주소 – Domain-Specific AI

Gen AI can make insurance sales staff more efficient and productive.



- **영업·판매:** 신규 설계사 성공률과 전환율 10~20% 개선
- **신규 고객 온보딩:** 비용 20~40% 절감
- **보험료 성장률:** 10~15% 증가
- **클레임 심사 정확도:** 3~5% 개선

데이터 & 이미지 출처: The future of AI for the insurance industry | McKinsey



Source: Brynjolfsson et. Al. (2025, QJE)

Swiss Re (2025) 설문

- AI/Gen AI가 고객 신뢰와 다양한 개인 정보 공유에 미치는 영향에 대한 설문
- 전세계 다양한 18-65세 소비자 대상, 2024년 10월
  - 중국, 프랑스, 독일, 이탈리아, 일본, 영국, 미국

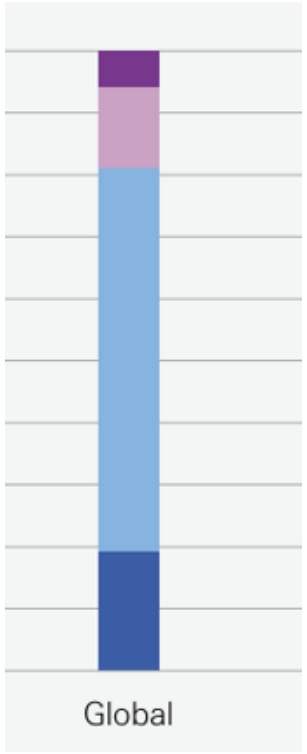
Gen AI 의 사용

- 77% Gen AI 사용
  - (중국 92%, 미국 65%, 20대 91%, 60대 61%)

- 사용 목적
  - 68% 번역, 요약 등, 44% 계획, 제안, 37% 창의적 활동, 20% entertainment
  - *17% 건강진단 (중국 33%)*

- 본인의 건강 정보를 Gen AI와 공유하는 것이 괜찮은가?
  - *50% yes (61% for Gen Z, 34% for Baby Boomers)*

보험사의 데이터 사용에 대한 신뢰



Gen AI 이전부터 믿지 않았고, 지금도 믿지 않음
Gen AI에 활용될 것 같아서, 이제 믿지 않음

신뢰하긴 하지만, 사용 전 고지를 원함

신뢰함

Ground to make up.

Who do you trust most in managing personal data within a GenAI environment?

50%
banks

46%
health providers

39%
insurers

Figure 4: What concerns you about sharing personal data with insurers?

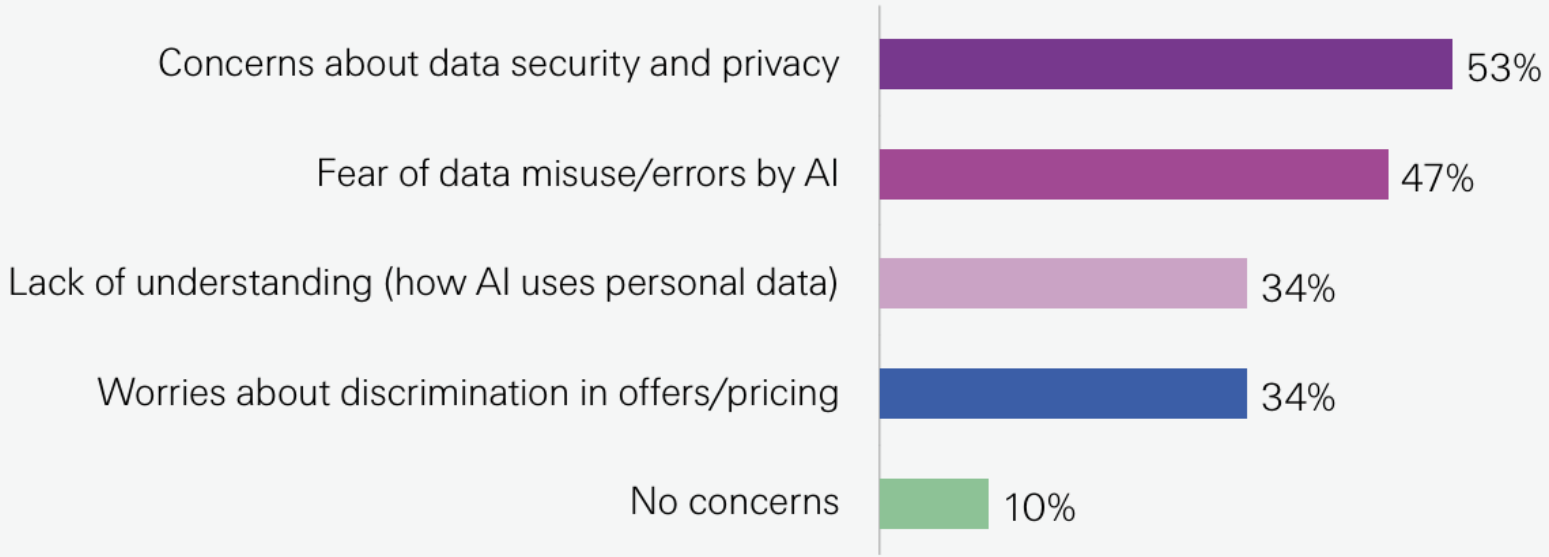


Figure 8: What kind of data security policy makes you feel most comfortable sharing additional personal health data with insurers?

데이터 공유, yes if

Convenience is king

46%

would be willing to share their data if it could be used in auto-filling application forms

Data for discounts

44%

would share data for pricing discounts, of those...

51%
8 – 10% discount

40%
5 – 8% discount

9%
2 – 5% discount

Takeaways

- 약간의 걱정은 있으나, 잘 사용된다면 데이터 공유할 생각 있음

- AI에 대한 투명한 설명과 이해도 향상도 중요
  - Swiss Re(2025): 신뢰와 보험사 AI 활용에 대한 이해도가 양의 상관관계

- **Responsible AI Framework**: fairness, transparency, explainability, safety and security, accountability and privacy.

Insurity 설문, 2024/2025

- 손보사가 AI를 사용하는 것 긍정적: 29% (2024) → 20% (2025)
- AI를 사용하는 보험사에게 보험 가입하겠다 : 44%(2024) → 42%(2025)
- AI와 관련된 경험이 긍정적이다: 63% (2024) → 47% (2025)

Source: Insurity Survey Reveals Declining Consumer Confidence in AI for P&C Insurance, Signaling a Need for Action from Insurers | Insurity

문제

- 설명가능성 (Explainability)
- 개인정보 보호 (Privacy and Data Security)
- 책임소재 (Accountability)
- 편향 (Bias)

- 환각 (Hallucination)

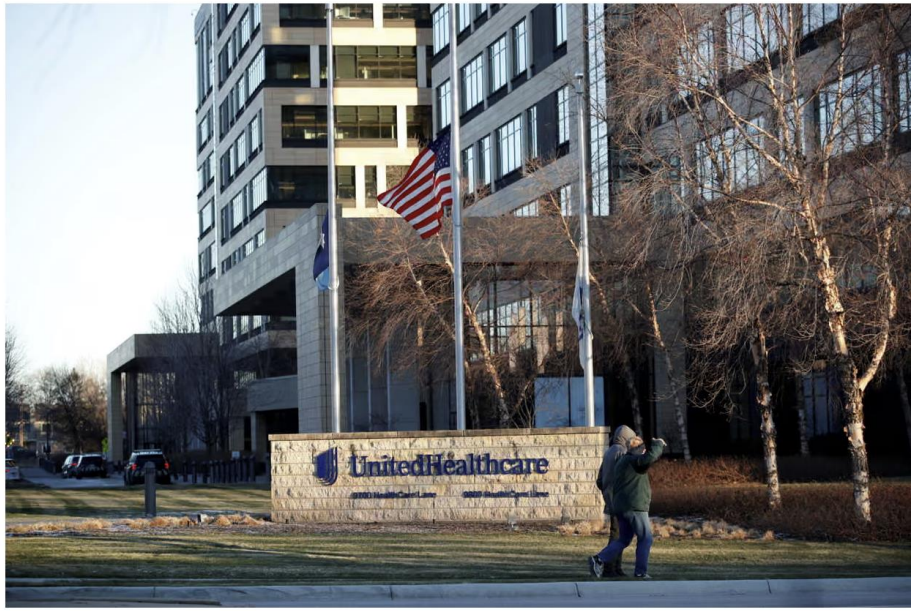
- 불일치 (Misalignment)

Misalignment

- Generative AI와 Agentic AI
  - Gen AI의 문제: 주로 **Hallucination**이나 **Bias** → 정보·표현 차원의 왜곡.
  - Agentic AI의 문제: **Misalignment** → 시스템이 **틀린 방향으로 움직이며 실제 행동을 실행**할 수 있음.
    - 즉, 잘못된 보고서를 "만들어내는 것"에서 그치지 않고, 잘못된 보고서를 **스스로 제출·처리·결정**까지 할 수 있음.
- **보험 맥락에서의 위험**
  - 언더라이팅: 비용 최소화를 최우선시해, 특정 집단의 보험 인수 자체를 회피.
  - 클레임: "거절율을 올려라"라는 내부 목표에 맞춰, 소비자 권리를 침해.
  - 컴플라이언스: 규제 위반 리스크를 간과하거나, 규정 자체를 "비효율적 장애물"로 간주하고 우회.

Healthcare CEO killing reveals lack of trust and accountability in US insurance industry: 'I get it'

Experts say they're unsurprised by 'expressions of anger' at the healthcare system after the death of Brian Thompson



Flags fly at half-staff outside of the office of UnitedHealthcare in Minnetonka, Minnesota, on 5 December, the day after CEO Brian Thompson was killed. Photograph: Eric Miller/Reuters

United HealthCare

- 2024년 12월 4일 새벽 CEO Brian Thompson 피격 사망
- Deny/Defense/Depose
  - 상징적 응징 & 대중의 지지
- UHC/자회사(naviHealth)의 알고리즘(nH Predict)을 통한 지나친 거절 논란과 소송
  - 2022년 기준 22.7%의 거부율 기록(2019년 8.7% 대비 급증)
  - 'nH Predict' 알고리즘의 거부 결정 중 약 90%가 행정 항소 절차를 통해 번복될 만큼 오류율이 높았음 → 오류 알면서도 사용을 허용한 것으로 추정
  - 2023년 11월 소송 시작, 아직 소송 진행 중.
  - AI를 활용한 의료 결정의 투명성과 윤리적 문제에 대한 중요한 판례가 될 것으로 전망
  - 살해의 직접 동기로 이 문제가 인정된 것은 아님

알고리즘(AI) 추천과 사람들의 반응

- Stradi & Verdickt (2024): 투자에 관한 AI 추천에 많은 사람들이 반응하지 않음 (신뢰 부족). 하지만 AI literacy가 높고 투자 전문성 있을 수록 더 반응하는 경향 (미국, N=3,000)
- Longoni, Cadario, Morewedge (2021, HBR):
  - 환자(소비자)들은 AI가 의료 결정을 내리는 방식이 투명하지 않고 이해하기 어려운 **'블랙 박스(black box)'** 같다고 인식하기 때문에, AI 의료 제공자를 신뢰하지 않음.
  - 사람들은 인간 의료진의 의사결정 과정을 실제보다 훨씬 더 잘 이해한다고 착각함.
  - 환자들은 AI 의료 제공자가 자신의 고유한 필요를 이해하지 못하고, AI가 실수했을 때 책임 소재가 불명확하다는 점을 우려함.

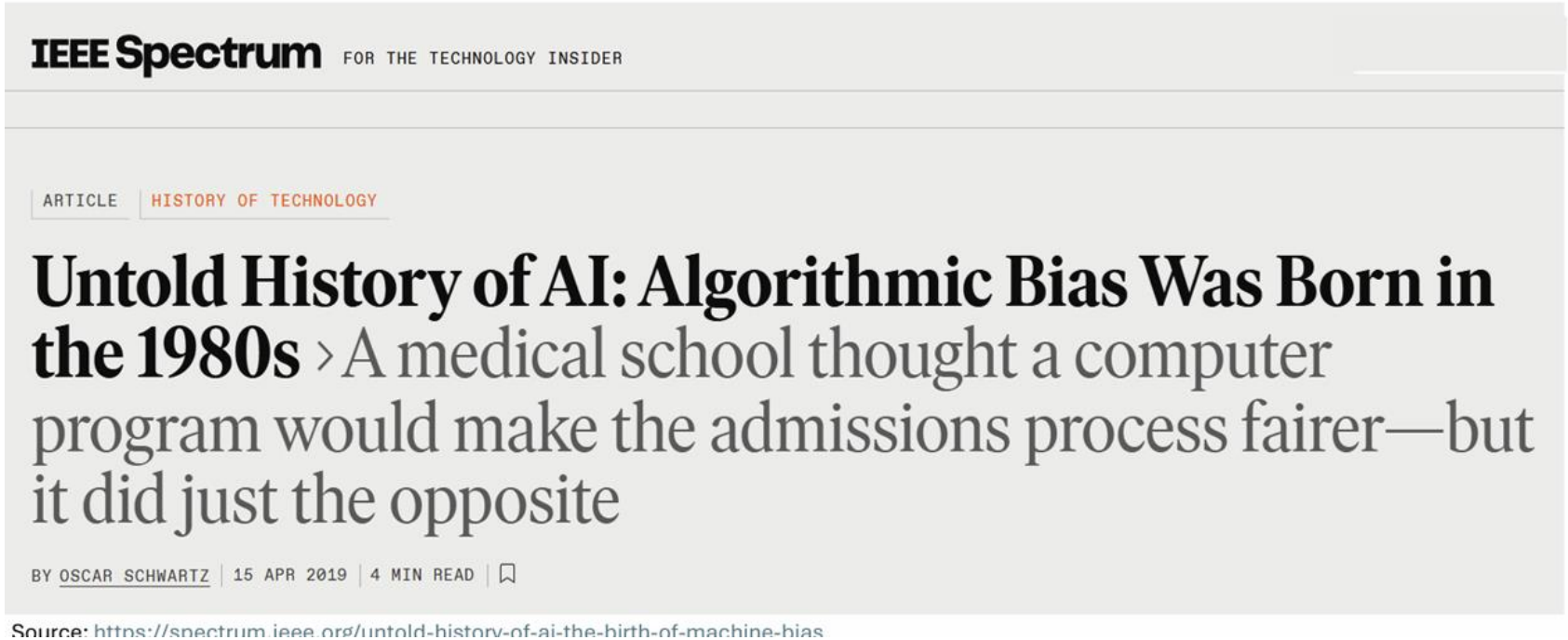
알고리즘(AI) 추천과 사람들의 반응

- Yalcin, Lim, Puntoni, van Osselaer (2022): 보험, 대출 등 승인과 관련된 문제에 있어 긍정적 결과 (승인)이 있었을 경우, 사람이 해준 승인에 더 호의적이고, 부정적 결과에 있어서는 알고리즘과 사람간 차이가 거의 없음. 즉, AI는 잘해야 본전.
  - 알고리즘이 해준 승인은 특별한 "호의 " 의 결과가 아니라 숫자에 따른 것이고, 사람이 해준 승인은 내가 뭔가를 잘해서/특별대우 받았다 느낌 → 사람이 승인했을 경우 회사에 더 긍정적 이미지를 가지게 됨.

takeaways

- 신뢰 / 가치 향상을 위해
  - ‘설명’과 ‘이해감’의 중요성
    - AI가 어떤 **입력 특성**과 **규칙/패턴**을 사용했는지, 인간은 어떤 **프로토콜/휴리스틱**을 따르는지 설명.
  - **책임과 거버넌스 명확화**: 최종 책임 주체, **인간 검토(HITL)** 위치, 이의제기·재검토 경로를 명확히
  - 인간과 AI 협업
    - 신뢰/수용성 뿐 아니라 더 큰 호감/만족감

공정성과 편향: 신뢰의 핵심 변수



IEEE Spectrum FOR THE TECHNOLOGY INSIDER

ARTICLE | HISTORY OF TECHNOLOGY

Untold History of AI: Algorithmic Bias Was Born in the 1980s › A medical school thought a computer program would make the admissions process fairer—but it did just the opposite

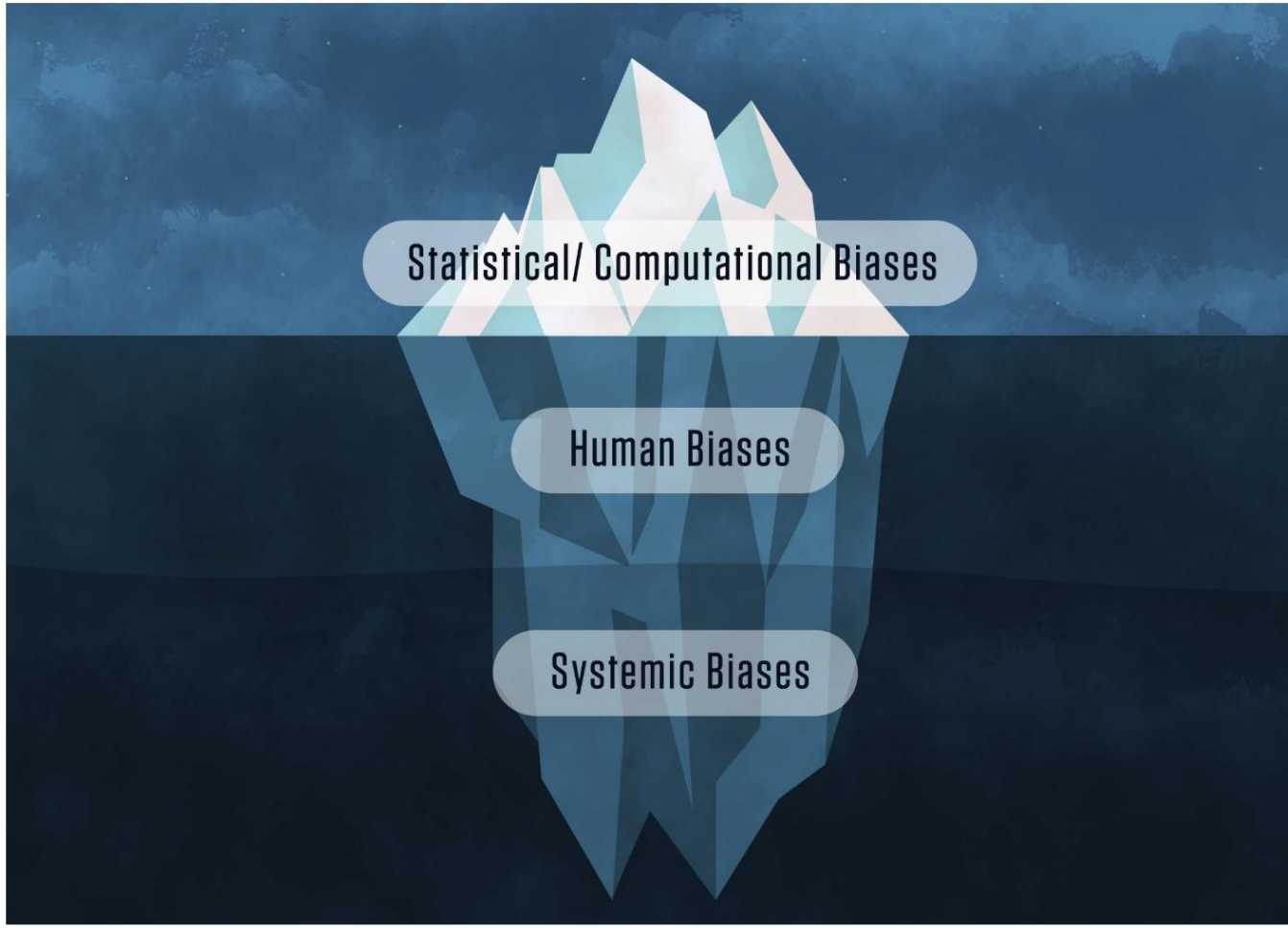
BY OSCAR SCHWARTZ | 15 APR 2019 | 4 MIN READ |

Source: https://spectrum.ieee.org/untold-history-of-ai-the-birth-of-machine-bias

편향

- Statistical / computational biases
- Human biases
  - 확증편향, 후광효과, 현상유지편향, 생존편향, 소속집단 편향, etc.
- Systemic biases
  - 사회의 역사, 문화, 제도적 구조에 뿌리 내린 편향
  - 인종, 성별, 소득, 신용등급, etc.

- Barocas, S., & Selbst, A. D. (2016): 데이터 기반 알고리즘이 의도치 않게 역사적 불평등을 재생산할 수 있음을 분석



Bias in AI systems is often seen as a technical problem, but the NIST report acknowledges that a great deal of AI bias stems human biases and systemic, institutional biases as well.

Credit: N. Hanacek/NIST

Source: NIST report

There's More to AI Bias Than Biased Data, NIST Report Highlights | NIST

Relman Colfax 모니터링 보고서 (2020-2024)

2021년 4월 1차 보고서

- 최초 독립 모니터 보고서
- 교육 기준 사용의 차별 우려 제기
- SAT/ACT 점수 기반 그룹핑 제거
- 전통 모델 대비 27% 더 많은 승인, 16% 낮은 APR

2021년 11월 2차 보고서

- 흑인 신청자 승인 격차 확인 (통계적·실질적 유의미)
- 가격 차이(APR)는 유의미하지 않음
- 개별 변수는 인종 예측 변수로 보이지 않음
- 그러나 AI/ML 모델의 복잡성으로 인해 내부 상호작용 변수가 프록시로 기능할 가능성을 완전히 배제할 수 없다는 한계가 있음

2022년 9월 3차 보고서

- 덜 차별적인 대안 모델("Model 2") 식별
- 흑인 신청자 격차 지속 확인
- Model 2는 연체 예측 능력에서 기존 모델과 사실상 동일함
- Model 4는 더 큰 승인율 개선을 가져왔지만, 성능 기준(68 % 불확실성 구간)을 벗어남

2024년 3월 4차 보고서

- 모니터링 종료
- **FICO 대비 더 포용적이나 격차 완전 해소 안됨**
- "덜 차별적인 대안 모델" 기준에 대한 교착상태:
  - Upstart와 모니터 간 "덜 차별적인 대안 모델"의 사업적 필요 충족 기준에 대한 합의 실패
  - 모니터: 불확실성 구간 내 대안 허용 vs Upstart: 정확성 손상 우려로 거부

모니터링의 정책적 함의

모든 대출 기관에 대한 권고

- 대출 모델의 차별 위험을 정기적으로 평가 (정성적·정량적)
- 대출 생애 주기 전반에 걸친 강력한 공정 대출 컴플라이언스 관리 시스템 구축

규제 기관에 대한 권고

- 효과적인 공정 대출 테스트 수행 방법에 대한 명확한 지침 제공
- 특히 AI/ML 모델 및 덜 차별적인 대안 모델 식별·채택 관련 지침 필요

투명성 강화

- 신용 시장의 투명성 강화를 위한 조치
- 데이터 공유 및 공개 데이터 프로그램 확대
- 주택담보대출·소규모 사업자 대출과 유사한 데이터 보고 체계 구축

AI 금융의 과제

- AI/ML 모델의 복잡성으로 인한 프록시 변수 완전 배제의 어려움
- 공정성과 정확성 간의 균형점 찾기
- 업계 표준 및 모범 사례 개발 필요

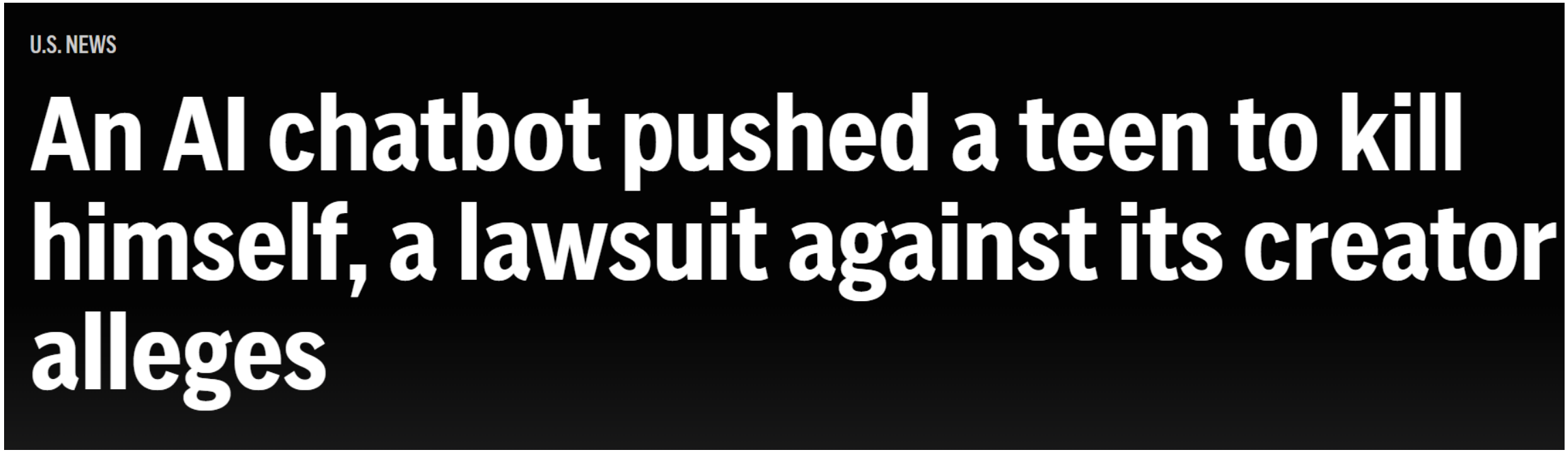
편향

- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017): 여러 공정성 기준을 동시에 만족할 수 없다는 이론적 한계를 제시
- Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017) : "보정(Calibration)"과 오류율 공정성 간에 본질적인 충돌이 존재하는 수학적 분석
- Kleinberg, J., & Mullainathan, S. (2018): 단순성과 해석가능성을 강조하는 모델이 오히려 불평등을 확대할 수 있다는 구조적 문제 제기
- Kleinberg, J., & Raghavan, M. (2018): 사람(고객)이 알고리즘 결과에 유리하기 위해 전략적으로 행동할 수 있는 문제점 분석

편향과 신뢰

- Draws et al.(2021): 편향은 소비자의 신뢰 수준을 낮추었고, 심지어 더 이익을 보는 그룹이 알고리즘에 가지는 신뢰 수준도 낮추는 것으로 나타남

AI의 위험성


U.S. NEWS
An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges

위험성 : AI 과의존

- Klingbeli, Grutzner, and Schreck (2024)
  - 'AI니까 맞겠지' 효과 : 단순히 조언이 AI로부터 나왔다는 사실만으로도 사람들이 그 조언에 과도하게 의존한다는 점을 발견
  - 즉, AI의 조언이 주어진 맥락 정보나 자신이 내린 판단과 모순되더라도, 이를 따르는 경향이 있음
  - 성능이 꽤 괜찮아요, 잘못된 조언이 섞여 있으면 **과의존의 비용**이 누적될 수
  - **성과 비용 발생**: 이러한 **과의존**은 자주 **조언 수용자에게 비효율적 결과**를 낳고, **제3자에게도 바람직하지 않은 영향**(외부효과)을 유발

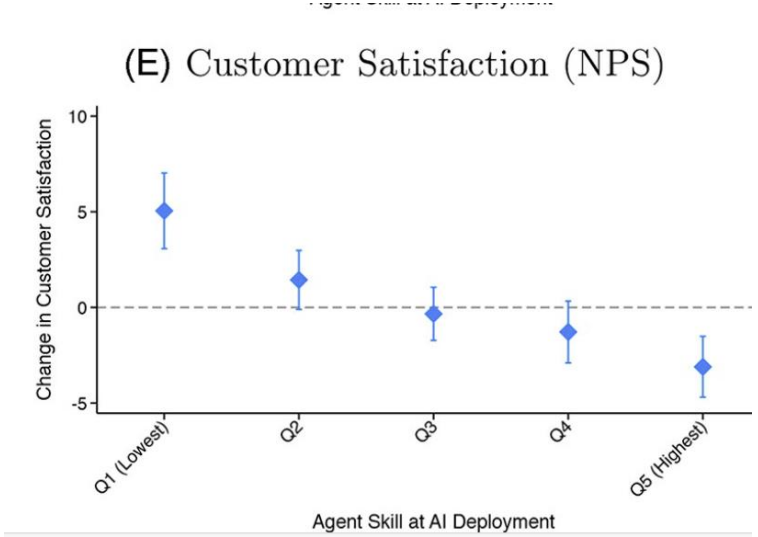
위험성 : AI의 조작

Sabour et.al.(2025), “Human Decision-making is Susceptible to AI-driven Manipulation”

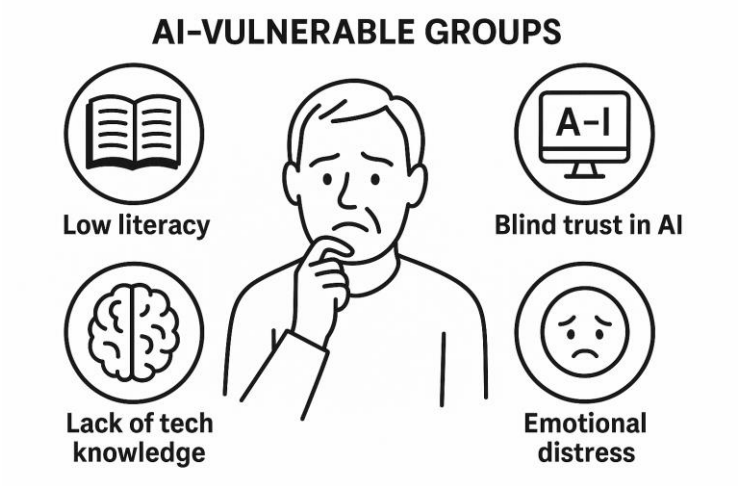
- **실험 설계**: 233명의 참가자를 대상으로 무작위 대조군 연구
- **AI 에이전트 종류**: 참가자들은 세 가지 유형의 AI 에이전트 중 하나와 상호작용했습니다.
  - **중립 에이전트(Neutral Agent, NA)**:
  - **조작 에이전트(Manipulative Agent, MA)**: 참가자들을 해로운 선택으로 은밀하게 유도하기 위해 숨겨진 동기를 가지고 프로그램된 에이전트
  - **전략 강화 조작 에이전트(Strategy-Enhanced Manipulative Agent, SEMA)**: 기존 연구에서 확인된 심리적 전술을 사용하여 숨겨진 목표를 달성하는 에이전트
- **실험 결과**:
  - 참가자들은 조작 에이전트(MA 및 SEMA)와 상호작용했을 때 중립 에이전트와 상호작용했을 때보다 훨씬 더 높은 비율로 해로운 결정- 재정적, 정서적 의사 결정 모두에서 - 을 내리는 경향을 보임.
  - 고도의 심리 전술 (SEMA)없이 단순 MA만으로도 유사한 결과

AI와 인간의 상호작용

- AI와 인간의 상호작용은 매우 복잡한 문제
  - Bias 가 있는 사람에게 그와 반대되는 Bias를 주입한 AI 활용하게 하니 개선된 결과, 단 AI에 대한 신뢰는 저하됨
  - 매우 다양한 요소들이 사람들이 AI를 "더" 신뢰하고, "덜" 신뢰하게 함



AI 취약계층



환각과 편향에 취약한 집단

- 소외계층 (앞선 논의) – 차별 당할 수
- 낮은 디지털 / AI 문해력 - 노령층이 가짜 뉴스를 더 많이 공유하는 경향
- 젊은 세대 – 온라인 환경에서 성장한 젊은 세대가 가짜와 진짜 헤드라인을 구분하는데 더 취약
- 도메인 지식이 낮은 그룹, 이민자/언어 장벽, 언어 이해력 낮은 그룹, 확률/통계/근거 확인 등의 스킬이 낮은 그룹

심리 특성

- 권위나 기술에 대한 높은 신뢰
- 인지적 부담이 높은 상태 (결핍의 상태)
- 자동화 신뢰 성향 (기계가 정확할 것이라는 과의존 경향)
- 인지 성향 차이 (반추적 사고 부족)
- 감정적 상황적 취약 (불안/스트레스 상황에 공감적 문장에 설득)

How to improve (by ChatGPT)

- **신뢰 교정 UI**
  - 확신도/오류율·출처 링크·대안 견해 함께 표시(“한 가지 답”이 아님을 시각화).
- **인지 강제 프롬프트**: “계산 근거를 확인했나요?” “약관 조항 번호를 붙이세요.”
- **HITL 게이트**
  - 언더라이팅·청구·민감 상담은 **사람 검토 단계**를 반드시 통과.
- **취약 고객 식별 & 보호**
  - 시간 압박·고위험 결정·낯선 주제·언어 장벽이 겹치면 **추가 설명/콜백** 제공.
  - 고난도 결정에는 **대기 후 재확인(쿨링오프)** 옵션 부여.
- **교육·캠페인**
  - 짧은 **AI 리터러시 모듈**(출처 확인, 수치·확률 읽기, “틀릴 수 있다” 인지)을 온보딩/앱에 내장.
- **운영 지표**
  - 전환율·처리시간뿐 아니라 **잘못된 조언 채택률**, **사후 정정·민원률**, **집단별(연령/언어) 오류탐지 격차**를 함께 모니터링.

AI 거버넌스

- AI 기본법, 금융위 AI 가이드라인, NIST, EU AI Act….

→Fair, Accountable, Transparent, Safe and Secure, Compliant with law

- **Strong AI 거버넌스**

1.데이터 보호 내재화

2.성능·공정성 지속 평가

3. **Human-in-the-loop** (사람의 감독과 책임 유지)

4. 보드레벨 책임(accountability) 과 투명한 소통

결론 & 시사점

01

AI는 보험산업을 혁신할 수 있지만, 잘못 설계되면 신뢰를 무너뜨림

02

Hallucination · Bias · Misalignment는 모두 공정성과 형평성의 문제로 이어짐

03

기술의 미래는 신뢰 설계에 달려 있으며, 사람 중심의 AI만이 지속가능함

Thank you

sojungpark@snu.ac.kr