SNU AI POLICY INITIATIVE

AI 윤리: 원칙에서 규범으로

거버넌스 원리로서의 AI 윤리

『인공지능 윤리: 원칙을 넘어 실천으로』
2021.2.17.

임 용, 이해성, 정종구
서울대학교 인공지능 정책 이니셔티브

http://sapi.co.kr

임 용
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#SAPI 서울대 인공지능 정책 이니셔티브

- **인공지능 관련 다양한 사회경제적·법적·정책적 이슈들을 연구 논의하기 위해 2017년에 발족**
  - 서울대학교 법과경제연구센터 산하 소셜 랩(Social Lab)
- **인공지능 정책 관련 우리나라의 주요 노드(node) 역할 수행**

#SAPI 서울대 인공지능 정책 이니셔티브

DAIG 매거진 창간호 (2020.12)

2020 Seoul AI Policy Conference (2020.9)

AI 윤리 1.0 – 지상(紙上)의 ‘원칙에서’

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#WhereWeAre AI 윤리: 우리의 현 주소

❑ 국가 기준도 마련되고…

사람이 중심이 되는
「인공지능(AI) 윤리기준」

2020. 12. 23

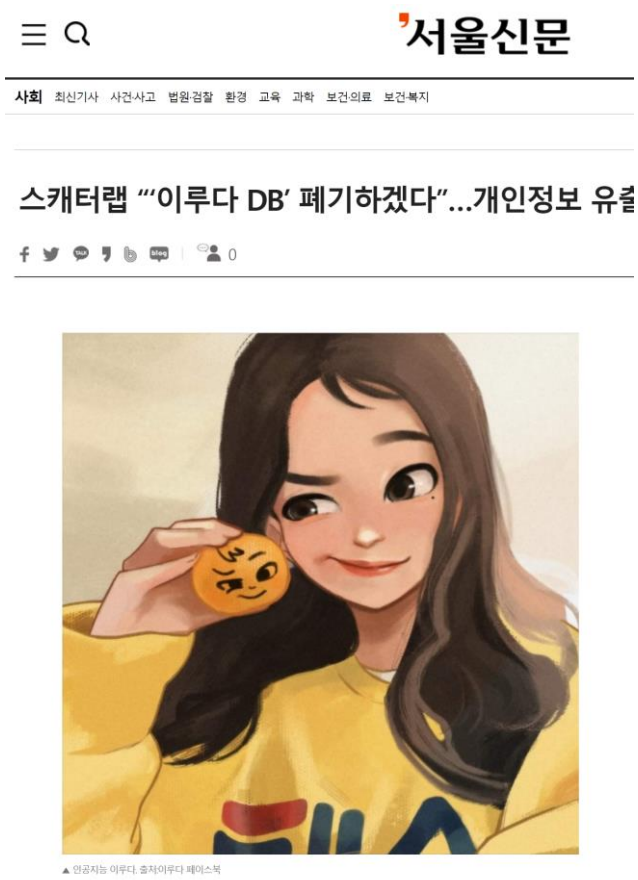
관계부처 합동

과기정통신부,
2020.12.23.자 보도자료(붙임),
https://www.msit.go.kr/SYNAP/skin/doc.html?fn=47eff97ca1b3a4bb17dd9f1fab86f9dd&rs=/SYNAP/sn3hcv/result/

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#WhereWeAre AI 윤리: 우리의 현 주소

- 경각심은 높아지고…

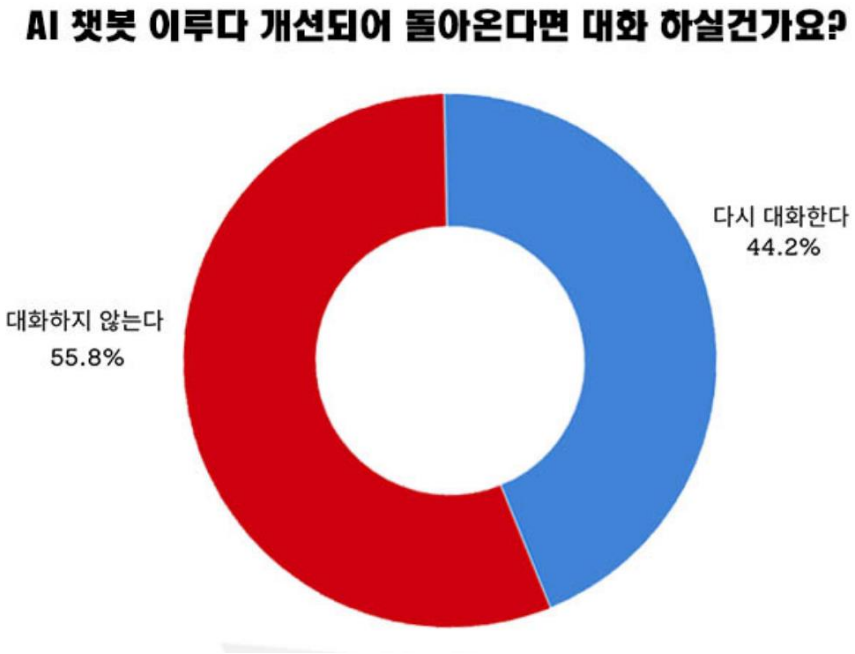


세계일보, 2021.1.15.자 기사, https://www.seoul.co.kr/news/newsView.php?id=20210115500080&wlog_tag3

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#WhereWeAre AI 윤리: 우리의 현 주소

- 시장의 불신은 쌓여가고…



FACTPL, '#25 AI 친구 루다, 설문 결과' (2021.1.15.),
https://stibee.com/api/v1.0/emails/share/Nh2MLLlYUvMcfoa-hq3S-0ReEEufNw==

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#WhereWeAre AI 윤리: 우리의 현 주소

❑ 촉구는 계속되고…

- "원칙 마련에서 행동 변환으로의 초점 전환이 다음 단계(shifting the focus from principle-formulation to translation into practice must be the next step)"

  - Jobin et. al., *Artificial Intelligence: the global landscape of ethics guidelines* (2019)

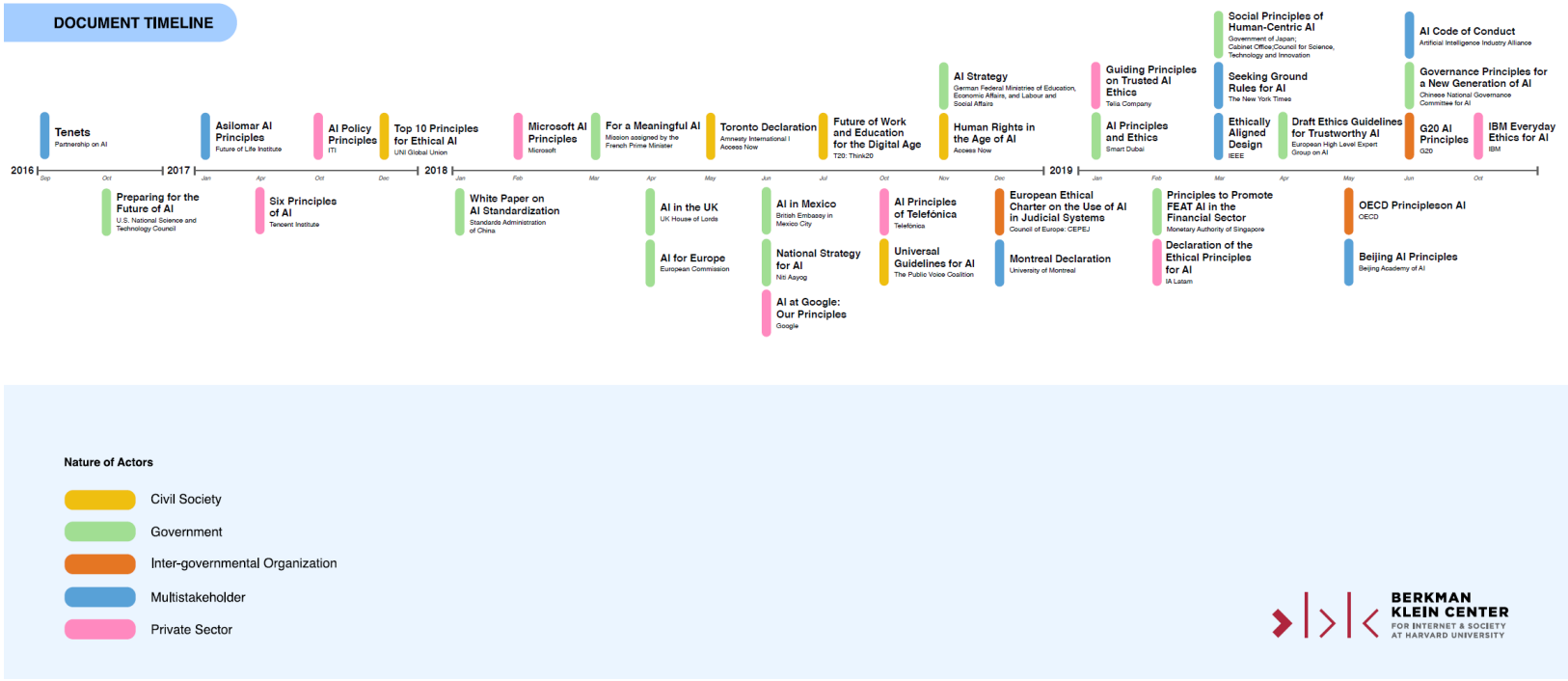
- "인공지능 개발자들이 시스템 이용자, 고객, 시민사회, 정부, 그리고 다른 이해관계자들로부터 인공지능을 책임있게 개발하고 있다는 신뢰를 얻기 위해, 원칙에 머물러 있지 말고 책임 있는 행동을 보여줄 메커니즘 마련에 포커스를 맞출 필요가 있다(In order for AI developers to earn trust from system users, customers, civil society, governments, and other stakeholders that they are building AI responsibly, there is a need to move beyond principles to a focus on mechanisms for demonstrating responsible behavior)"

  - Brundage et. al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims* (2020. 4.)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LookingBack AI 윤리: 어디까지 왔는가?

❑ 원칙 모색 중심의 논의 전개

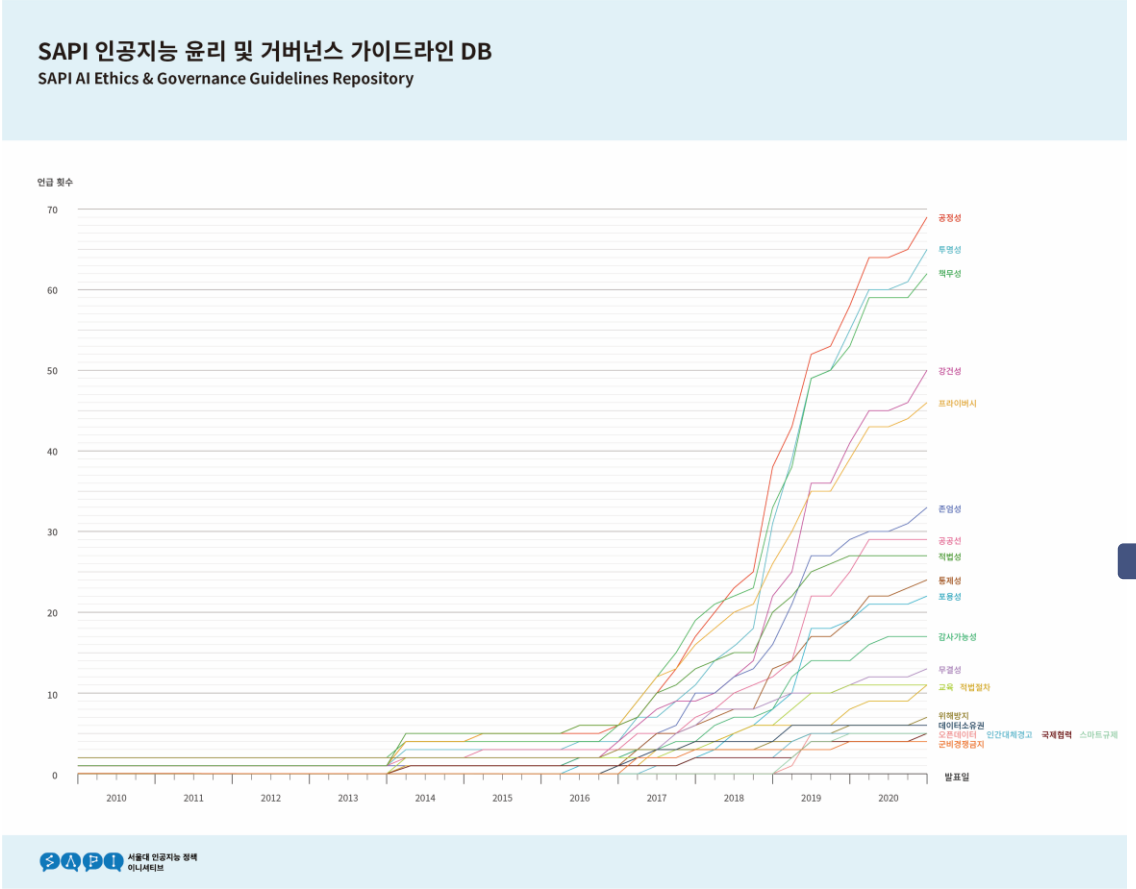


Harvard Berkman Klein Center,
Principled Artificial Intelligence (2020.1)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LookingBack AI 윤리: 어디까지 왔는가?

❑ 다양하게 제시된 윤리 원칙



임용, 이해성, 정종구,
「인공지능 거버넌스: 윤리와 규범 사이에서」
(*forthcoming* 2021)

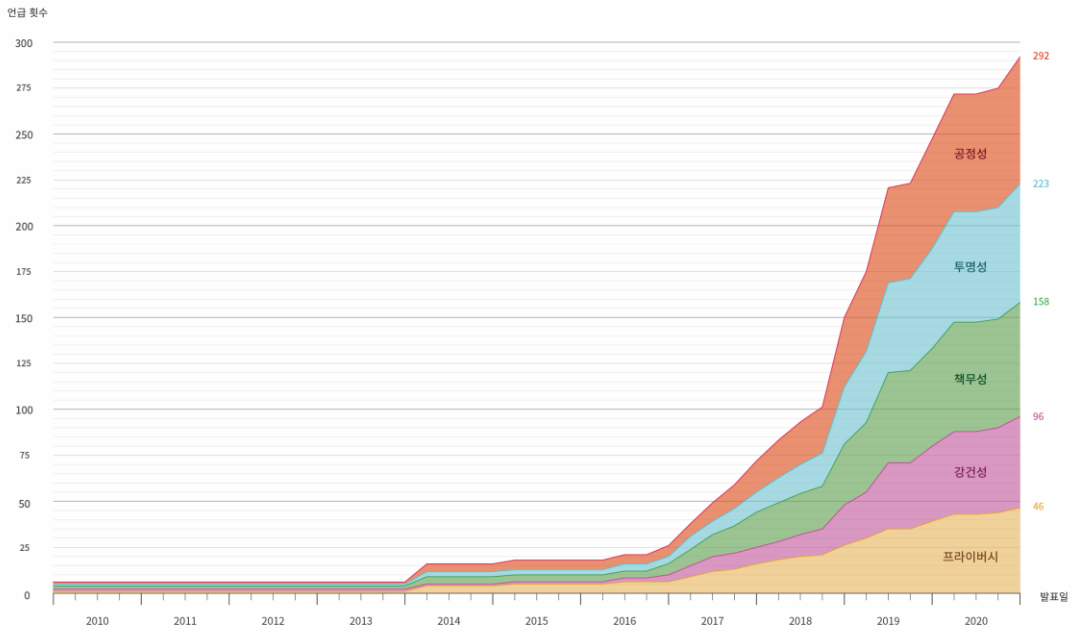
임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LookingBack AI 윤리: 어디까지 왔는가?

❑ 수렴되는 (듯한) 모습…

- 공정성(Fairness, Non-discrimination)
- 투명성(Transparency, Explainability)
- 책무성(Accountability)
- 강건성(Safety, Security)
- 프라이버시(Privacy)
- …

SAPI 인공지능 윤리 및 거버넌스 가이드라인 DB
SAPI AI Ethics & Governance Guidelines Repository



SAPI 서울대 인공지능 정책 이니셔티브

임용, 이해성, 정종구,
「인공지능 거버넌스: 윤리와 규범 사이에서」
(*forthcoming* 2021)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#WhyPrinciples AI 윤리(원칙)는 필요한가?

❑ Principlism 기반 접근에 대한 우려

- AI 분야와 의료 분야간의 차이(Mittelstadt 2019)

❑ AI 분야에서의 윤리 기반 접근의 효용

- 사회적 요구에 대한 대응(Accenture 2019)
- 인공지능 기반 사회로의 전환이라는 공동 과제의 논의의 장(Gasser & Almeida 2017: “shared responsibility”)
- 법(규제)과 기술(시장)의 시간적 격차로 인한 문제 해결(Larsson 2020)
- AI의 과다 이용과 과소 이용의 동시 회피 수단(Floridi et. al. 2018: “dual advantage”)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#WhyPrinciples AI 윤리(원칙)는 필요한가?

❑ AI 분야에서의 윤리 '원칙' 논의의 가치 – '소통'

- AI 윤리(원칙)는 급속한 기술 변화와 패러다임 전환으로 인한 불확실성에 직면한 다양한 사회적 주체들간 소통 수단으로서 기능(Morley et. al. 2019: "common language")
  - 소통은 계속 진행 중

AI의 윤리적 거버넌스의 5개 기둥(Winfield & Jirotka 2018)

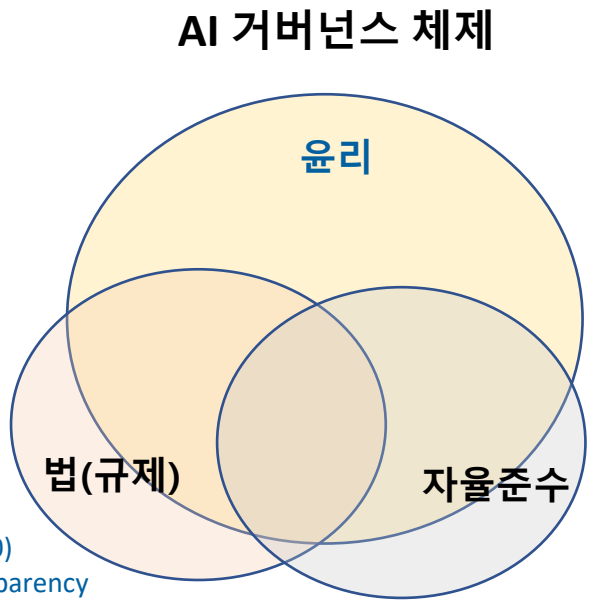
- 윤리적 행위 준칙의 발표 → AI 윤리 원칙의 마련은 소통의 시작
- 구성원 전원의 윤리 및 책임있는 혁신 교육/훈련 수행
- 책임있는 혁신의 실천
- 윤리적 거버넌스에 관한 투명성 유지
- 윤리적 거버넌스의 가치 존중

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LessonsLearned 우리는 AI 윤리 논의를 통해 무엇을 배웠는가?

❑ #Lesson1 – 윤리는 AI 거버넌스의 한 축(일 뿐)이다

- 로보틱스(Robotics) 분야와의 비교: Ethics-Standards-Regulation (Winfield & Jirotka 2018)



(예) EU P2B Regulation (2020)
Guidelines on Ranking Transparency
기존 법령

(예) AI Blindspot Cards (2019)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LessonsLearned 우리는 AI 윤리 논의를 통해 무엇을 배웠는가?

❑ #Lesson1 – 윤리는 AI 거버넌스의 한 축(일 뿐)이다

- 논의를 '윤리'에서 '거버넌스'로 확대할 필요: 단순한 레이블(labelling)의 문제만은 아님
  - '윤리'의 관점에서 포섭이 쉽지 않은 원칙 존재(설명가능성 등)
  - "Ethics washing" 방지 필요(Schiff et. al. 2020 外)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LessonsLearned 우리는 AI 윤리 논의를 통해 무엇을 배웠는가?

❑ #Lesson2 – 인간과 AI간 상호작용(Human-AI Interaction)에 주목할 필요가 있다

- Human-AI Interaction(HCI)이 AI 관련 리스크의 예측, 발견 및 해결에 중요한 영향 미침(Zhou et. al. 2019)

Xiaoice, https://techcrunch.com/2020/07/12/microsoft-spins-out-5-year-old-chinese-chatbot-xiaoice/

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LessonsLearned 우리는 AI 윤리 논의를 통해 무엇을 배웠는가?

❑ #Lesson3 – 논의가 현장(현실)과 연결되어야 한다

- COMPAS 사건: Minority Reort의 전조?

Angela Christin, *The Mistrials of Algorithmic Sentencing*, https://logicmag.io/justice/the-mistrials-of-algorithmic-sentencing/

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LessonsLearned 우리는 AI 윤리 논의를 통해 무엇을 배웠는가?

❑ #Lesson4 – 인간의 역할(Human in the Loop)이 중요하다

- 적절한 개입을 할 수 있는 지식, 경험 및 전문성을 갖춘 인력의 양성 필요

MIT Technology Review,
This is the Stanford vaccine algorithm that left out frontline doctors (2020.12.21.)
https://www.technologyreview.com/2020/12/21/1015303/stanford-vaccine-algorithm/

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LessonsLearned 우리는 AI 윤리 논의를 통해 무엇을 배웠는가?

- **#Lesson5 – 윤리적(거버넌스) 문제의 개선(완화)는 가능하고…**
  - 긍정적이고 윤리적인 기능의 강화와 개선 가능(Morley et. al. 2019: “progressive increase”)

MIT Technology Review,
A biased medical algorithm favored white people for health-care programs (2019.10.25.)
https://www.technologyreview.com/2019/10/25/132184/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LessonsLearned 우리는 AI 윤리 논의를 통해 무엇을 배웠는가?

❑ #Lesson6 – …원칙간 '트레이드 오프(Tradeoff)'는 피할 수 없다

- 가치와 원칙간의 상충 가능성과 형량의 요구(Goodman & Flaxman 2017 外)

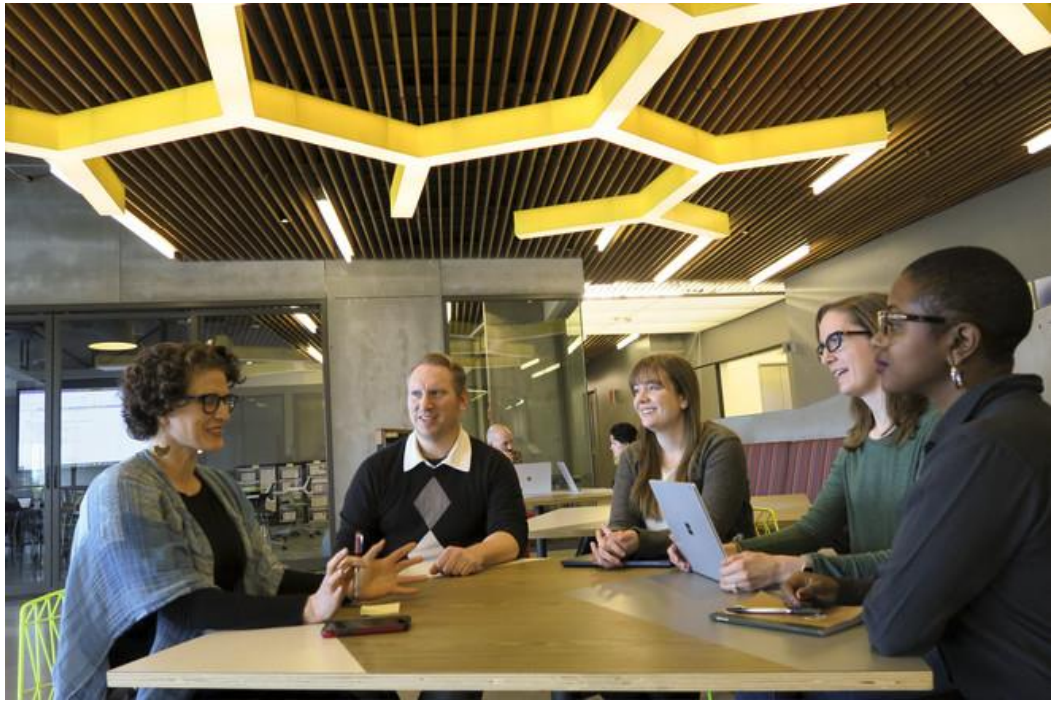
BBC,
German parents told to destroy Cayla dolls over hacking fears (2017.2.17.)
https://www.bbc.com/news/world-europe-39002142

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LessonsLearned 우리는 AI 윤리 논의를 통해 무엇을 배웠는가?

❑ #Lesson7 – 앞으로 다양성(diversity)은 선택이 아니라 필수다

- 데이터 수집부터 조직 구성과 의사결정 과정 전반에 걸쳐 요구(다양한 경험과 시각 포섭)



WSJ,
A Crucial Step for Averting AI Disasters
(2019.2.13.)
https://www.wsj.com/articles/a-crucial-step-for-avoiding-ai-disasters-11550069865

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LessonsLearned 우리는 AI 윤리 논의를 통해 무엇을 배웠는가?

❑ #Lesson7 – 앞으로 다양성(diversity)은 선택이 아니라 필수다

- 인공지능 비서의 개발 사례

"I'm gay."

"I'm AI."

"Cool. I'm AI."

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#LessonsLearned 우리는 AI 윤리 논의를 통해 무엇을 배웠는가?

❑ #Lesson8 – (잠정적) '사용 배제'도 거버넌스 방안이 될 수 있다

- 측정하기 어려운 높은 리스크 포착 또는 해결하기 어려운 이해관계의 상충 직면 등에서 고려 가능(WEF 2019)

Minneapolis poised to ban facial recognition for police use

Committee voted 12-0 in favor of ban, advancing it to city council, months after actions of city's police sparked racial reckoning in US

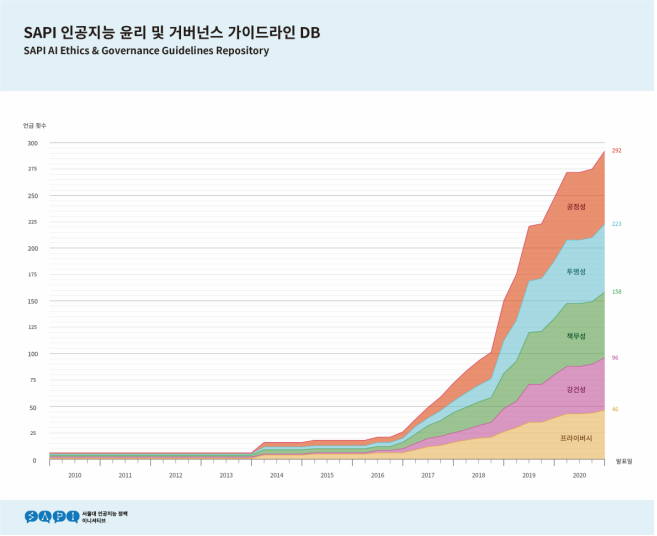
The Guardian, *Minneapolis poised to ban facial recognition for police use* (2021.2.12.) https://www.theguardian.com/us-news/2021/feb/12/minneapolis-police-facial-recognition-software

AI 윤리 2.0 – 현장의 '규범으로'

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#PraticablePrinciples 기업 현장의 실천적 규범으로서의 윤리 원칙

- **현장에서 규범화의 선두에는…**
  - 프라이버시
  - 투명성
  - 강건성
  - 공정성
- **물론 다른 원칙(존엄성 등)의 경시는 피해야…**
  - Non-maleficence v. Beneficence

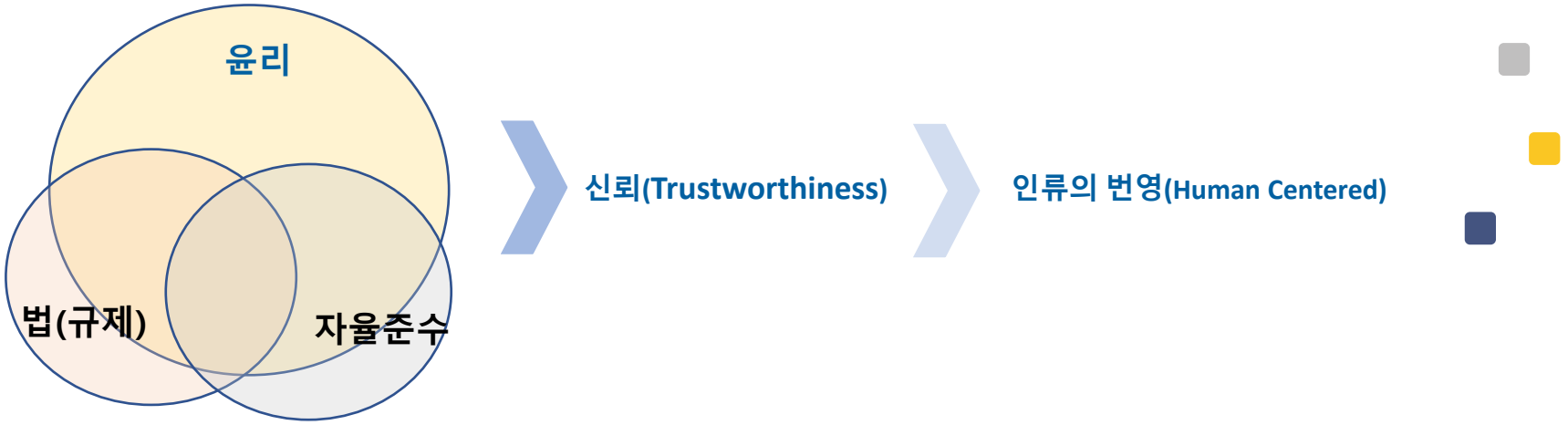


임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ **#VigilantlyVerifiedTrust** AI 거버넌스의 목표는 '지속적 견제를 통해 검증된' 신뢰(Vigilantly Verified Trustworthiness)의 확보다

AI 거버넌스 체제

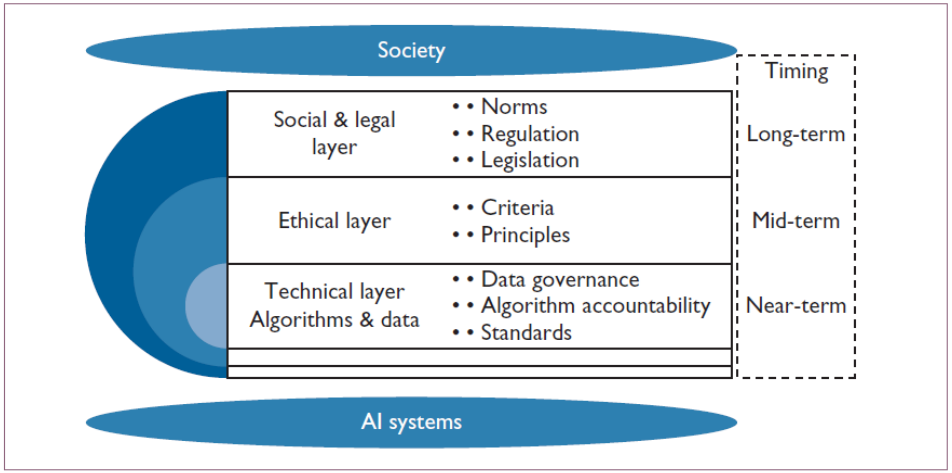


임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ #SocioLegalGovernance AI 거버넌스 체제의 사회·법적(socio-legal) 측면에 관심을 가질 시점이다

- 목표는 '도덕적'인 'AI 기술'의 개발이 아니라, '신뢰 가능'한 AI 기반 '사회 시스템'의 조성임
  - 조직의 윤리 문제: Professional ethics → Organizational ethics (Mittelstadt 2019)



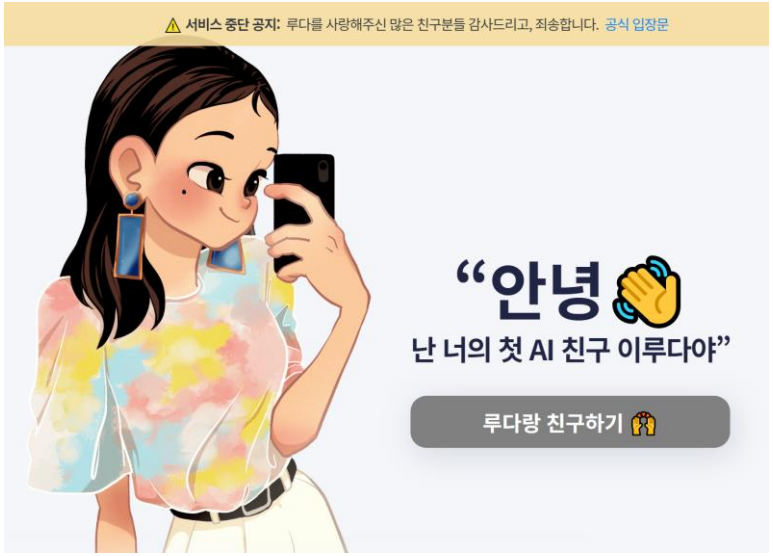
Urs Gasser & Virgilio A.F. Almeida, *A Layered Model for AI Governance,* 21(6) IEEE Internet Computing 60 (2017)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ #EthicsAsProcess 윤리도 프로세스(process)의 관점에서 접근할 필요가 있다

- 프로세스를 통해 구현하는 윤리적 거버넌스 정착 필요
  - Luda.ai 사례를 통해 보는 프로세스의 중요성



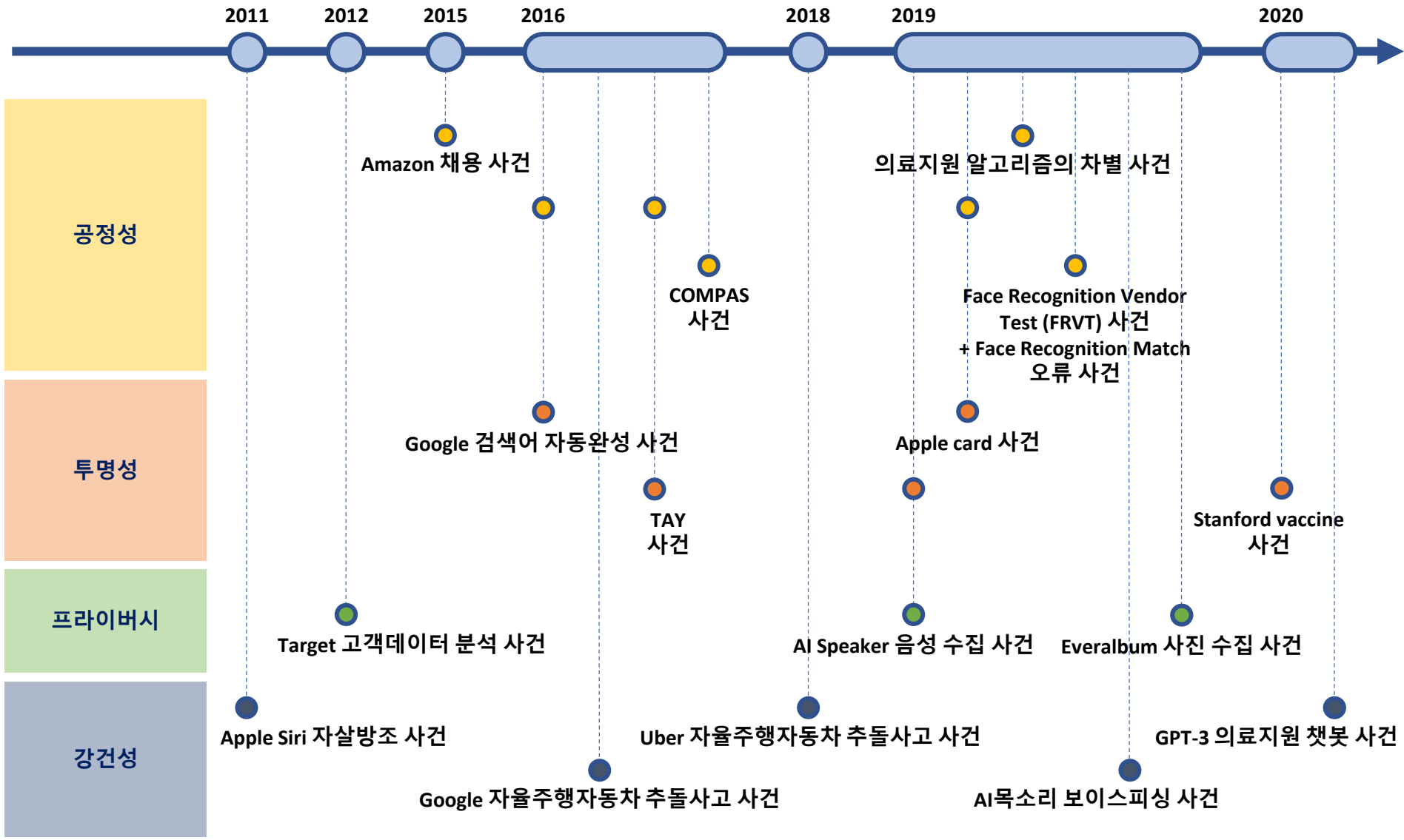
임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ #EthicsAsProcess – 윤리도 프로세스(process)의 관점에서 접근할 필요가 있다

- Luda.ai 사례를 통해 (재)노출된 문제들
  - 개인정보의 무단 이용(논란) → 프라이버시, 투명성, 책임성
  - 차별적 · 혐오발언 → 공정성, 안전성
  - 인공지능의 젠더링과 성적 대상화 → 공정성, 안전성
- 그런데, 실은 그 전에도 사건 사고들은 계속 있어왔고…

#EthicsAsProcess AI 관련 사건 사고(미국·유럽)



SAPI 동계 단기 프로젝트 연구진(2021, DRAFT)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ #EthicsAsProcess – 윤리도 프로세스(process)의 관점에서 접근할 필요가 있다

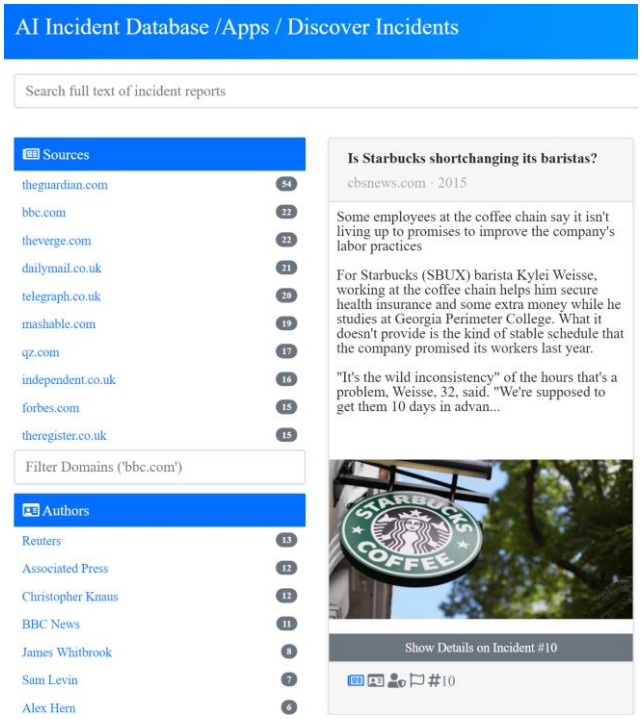
- Luda.ai의 개발·운용 단계에서 결여되었던 것 한 가지 – **신뢰 확보가 가능한 프로세스**
  - 인공지능이 의사결정의 기반이 될 경우 기술적인 완결성(목적 달성)만으로는 신뢰 확보 어려움 – '사회적인 가치의 적정한 고려'와 그에 대한 '상대방의 적절한 인지' 둘 다 필요
    - (일차적으로는) '**Embedded Diversity**'를 통한 해결: 문제되는 인공지능 기반 제품·서비스와 관련되는 사회적 가치들이 planning-building-deploying의 단계 전반에 걸쳐 적절히 고려, 반영 및 검증될 수 있도록 다양성이 내제된 (인적·물적) 조직 및 프로세스를 구축
  - 그런데, 사회적 가치의 불명확성과 불확정성(상황 구체성 포함)으로 인한 어려움 존재

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ #AcceptableAI – '적정한(수용가능한)' AI의 구현을 생각할 때다

- AI와 관련된 오류는 필연
  - Model decay
  - Complexity
  - Probabilistic nature



PAI AI Incident Database
https://incidentdatabase.ai/

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ #AcceptableAI – '적정한(수용가능한)' AI의 구현을 생각할 때다

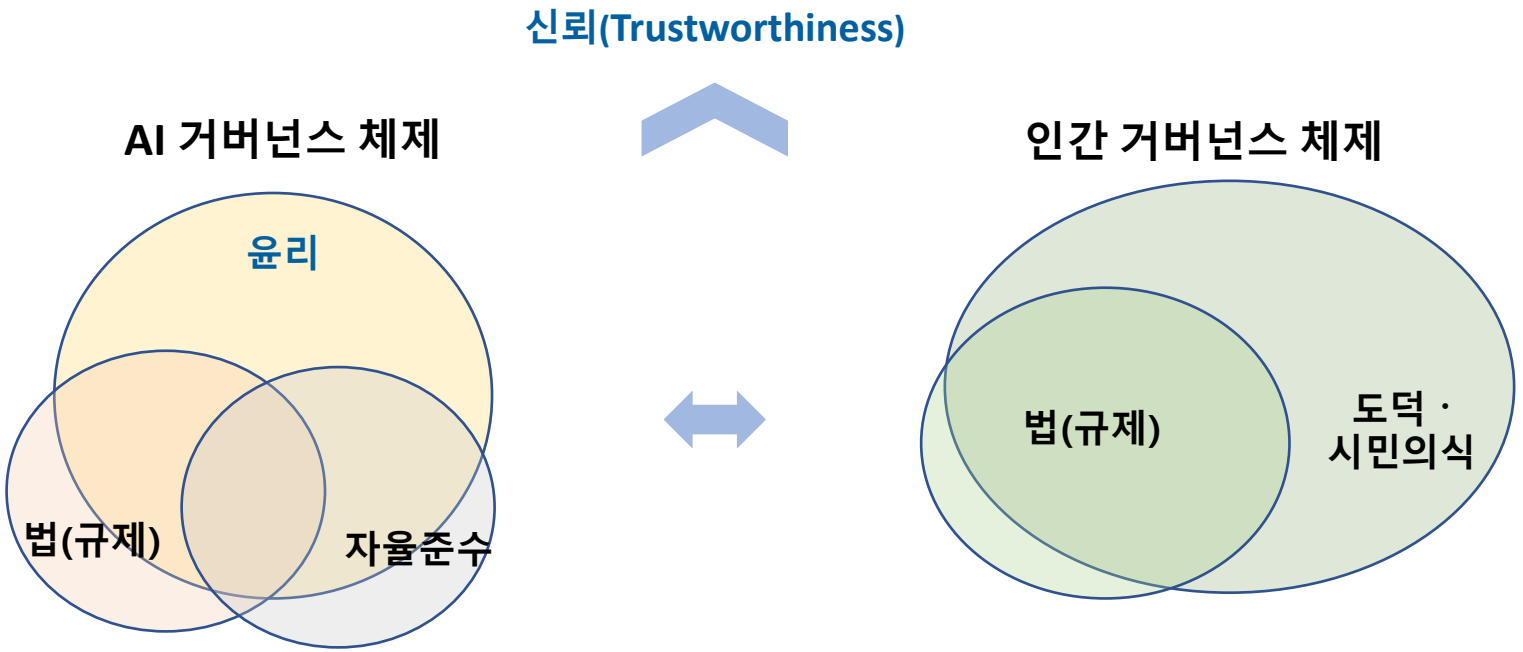
- 오류의 허용 여부 및 수인 가능한 오류율 논의 필요
  - 구체적인 상황마다, 이용자 혹은 상대방마다, 이용 목적별로 오류의 허용 여부와 사회적으로 수인 가능한 오류율에 차이 존재: 규범적 판단 필요
  - *cf)* 경쟁법의 '유효경쟁(Workable competition)'
- 오류의 허부 내지 오류율 등에 대한 합의가 어려울 경우 민주적 절차를 통한 해결이 필요할지도
- 그런데, 결과 기반의 신뢰 확보(outcome driven trustworthiness)도 한계가…
  - 상황마다 적합한 허용 오류율의 결정을 통해서 모든 문제를 해결하기는 어려움(신뢰 확보 실패)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ #AcceptableAI – '적정한(수용가능한)' AI의 구현을 생각할 때다

- Acceptable AI 달성을 위한 시스템 구축: 인간 기반의 의사결정 v. 인공지능 기반의 의사결정



임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ #AcceptableAI – '적정한(수용가능한)' AI의 구현을 생각할 때다

- Acceptable AI 달성을 위한 시스템 구축: 인간 기반의 의사결정 v. 인공지능 기반의 의사결정
  - (질문) 우리 사회가 인간 기반의 의사결정을 신뢰(수용)하는 이유는 무엇인가?

[Case] 인간 판사의 결정: 판사의 훌륭한 인품(도덕성)보다는...

- 자격(변호사시험)
- 신분보장(탄핵)
- 판결문의 공개
- 절차적 보장
- 상소의 가능성

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ #AcceptableAI – '적정한(수용가능한)' AI의 구현을 생각할 때다

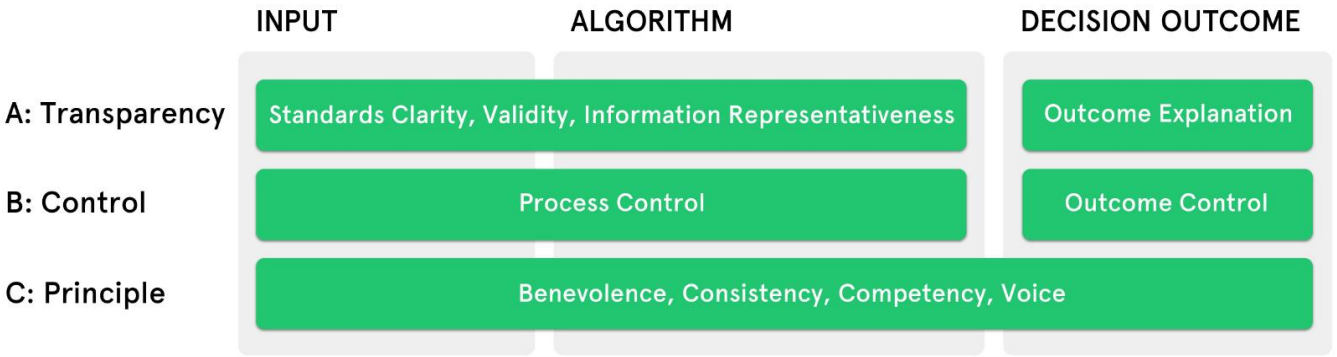
- Acceptable AI 달성을 위한 시스템 구축
  - 코드, 조직, 제품 개발 및 의사결정 과정 등 전반에 걸쳐 시스템적으로 '견제와 균형(check and balances)'의 구현 모색(MS 2020: "good tension")
    - Meta-AI
    - Ethics board
    - External Audit
  - 이와 같은 시스템을 통과한 AI 제품/서비스의 경우 적법성 판단에 고려(*cf*. business judgement rule)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#RulesOnTheGround 현장에서 실천하는 AI 윤리

❑ #AcceptableAI – '적정한(수용가능한)' AI의 구현을 생각할 때다

- 신뢰 확보를 위한 절차적 공정성(Procedural Fairness)에 대한 관심 고조
  - 지금까지 결과적(분배적) 공정의 기술적 구현에 집중되었던 논의가 인식(perception) 기반의 절차적 공정을 통한 신뢰 확보 문제로 확장되고 있음



Lee et. al., *Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation* (2019)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#WorkToDo 향후의 과제(예시)

❑ 원칙과 툴의 매칭 작업: 학제간 융합적 협업 필요

	Business and use-case development Problem/improvements are defined and use of AI is proposed	**Design Phase** The business case is turned into design requirements for engineers	**Training and test data procurement** Initial data sets are obtained to train and test the model	**Building** AI application is built	**Testing** The system is tested	**Deployment** When the AI system goes live	**Monitoring** Performance of the system is assessed
Beneficence							
Non-maleficence	(Cavoukian et al. 2010) outline 7 foundational principles for Privacy by Design: 1.Proactive not reactive: preventative not reactive. 2.Privacy as the default 3. Privacy embedded into design 4. Full functionality = positive sum, not zero sum 5. End-to-end lifecycle protection 6. Visibility and Transparency 7. Respect for user privacy	(Oetzel and Spiekermann 2014) set out a step-by-step privacy impact assessment (PIA) to enable companies to achieve 'privacy-by-design'	(Antignac et al. 2016) provide the python code to create *DataMin* (a data minimiser—a pre-processor modifying the input of data to ensure only the data needed are available to the program) as a series of Java source code files which can be run on the data sources points before disclosing the data.	(Kolter and Madry 2018) provide a practical introduction, from a mathematical and coding perspective, to the topic of adversarial robustness with the idea being that it is possible to train deep learning classifiers to be resistant to adversarial attacks: https://adversarial-ml-tutorial.org/	(Dennis et al. 2016) outline a methodology for verifying the decision-making of an autonomous agent to confirm that the controlling agent never deliberately makes a choice it believes to be unsafe	(*AI Now Institute Algorithmic Accountability Policy Toolkit*) provides a list of questions policy and legal advocates will want to ask when considering introducing an automated system into a public service and provides detailed guidance on where in the procurement process to ask questions about accountability and potential harm https://ainowinstitute.org/aap-toolkit.pdf	(Makri and Lambrinoudakis 2015)outline a structured privacy audit procedure based on the most widely adopted privacy principles: -Purpose specification -Collection limitation -Data quality -Use retention and disclosure limitation -Safety safeguards -Openness -Individual participation -Accountability
Autonomy							
Justice							
Explicability							

Morley et. al., *Applied AI Ethics typology, What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices* (2019)

임 용, 이해성, 정종구
서울대학교 법학전문대학원
서울대학교 인공지능 정책 이니셔티브

#WorkToDo 향후의 과제(예시)

- 원칙과 툴의 매칭 작업: 학제간 융합적 협업 필요
- Microethics: “moral-semantic trilemma”의 극복
- Virtue Ethics: “과학 분야로서의 윤리(Ethics as a Scientific Discipline)”
- AI 오류의 대응/대책 시스템 구축
- 법적 책임의 규명과 분배 문제
- ….

질의·응답

감사합니다