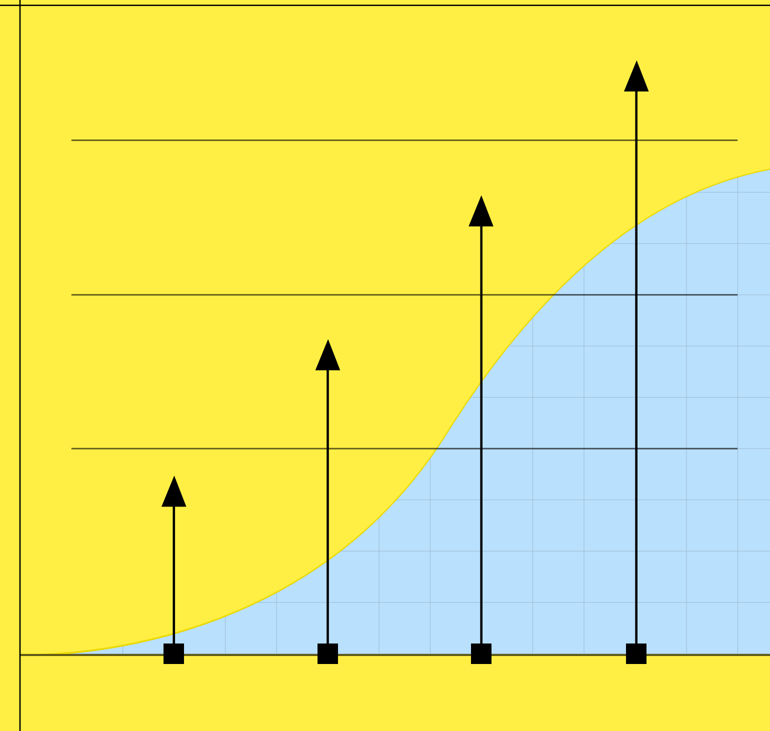
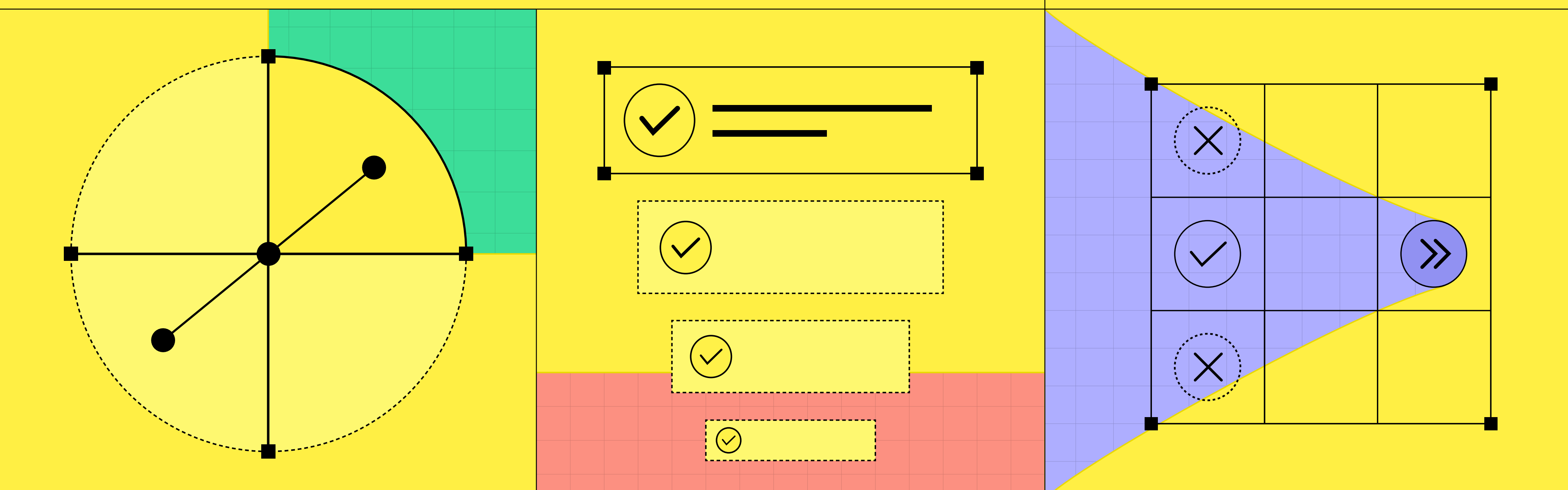
책임 있는 AI를 위한 생애주기별 자가점검 도구 설명서

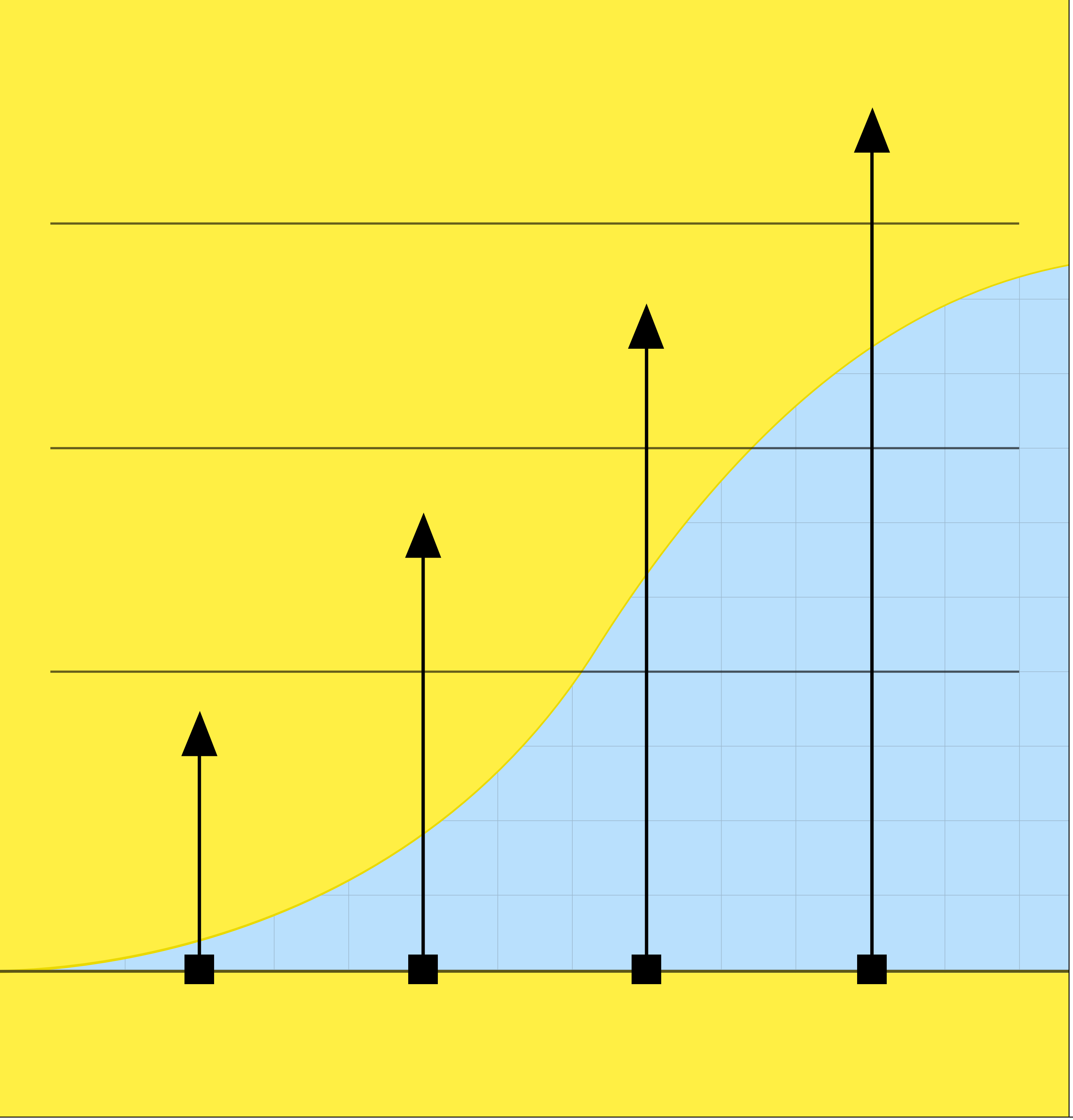




발행일자
2024.07.31

kakao

목차



여는 글: 카카오의 AI 윤리 거버넌스 고도화를 향한 과정

1

1 여는 글: 카카오의 AI 윤리 거버넌스 고도화를 향한 과정

인공지능(Artificial Intelligence, AI)은 기술을 통해 인간의 지각, 학습, 추론 등의 능력을 구현함으로써 무한한 가능성을 제시하는 동시에 **AI 기술에 대한 사회적 책임**이 무엇이고 그 **책임을 어떻게 실천**할 수 있을지에 대한 화두를 던집니다.

카카오는 꾸준히 **카카오만의 책임 있는 AI 실행 방안**을 고민해 왔습니다. 그간 **AI 윤리**[01] **정책을 수립**하고, 담당 조직을 발족함으로써 **AI 윤리 거버넌스 체계 구축**의 기반을 닦고, 관련 **정책을 실천**하기 위해 AI 윤리 자가점검 도구를 개발하는 등 지금도 **카카오는 AI 기술의 현장에서 AI 윤리를 내재화**하기 위해 노력하고 있습니다.

이러한 **카카오의 AI 윤리 거버넌스 고도화를 향한 과정**은 앞으로도 계속될 것입니다. 단기적으로 결과물을 만들어내는 것에 국한되기보다 **더 넓은 관점에서 책임 있는 AI를 위한 비전을 세우고, 카카오와 주요 계열사가 그 비전을 향해 함께 나아갈 수 있도록 하는 것**이 이 과정의 핵심 가치라 할 수 있습니다.

1 여는 글: 카카오의 AI 윤리 거버넌스 고도화를 향한 과정

정책 개선 그리고 실천의 영역으로

2018년 1월 **카카오 알고리즘 윤리헌장**이 공개되었습니다. 이후, 카카오는 포용성(2019년), 아동과 청소년 보호(2022년) 및 프라이버시(2022년) 관련 조항을 추가함으로써 알고리즘 윤리헌장을 꾸준히 개정해 왔습니다. 또한, 알고리즘 윤리헌장이라는 선제적 **정책 선언을 실천의 영역으로** 가져가기 위해 증오발언 근절 원칙을 이모티콘 제작 가이드 등 자사 서비스에 적용하였으며(사회적 차별에 대한 경계), 다음 뉴스 배열 설명서를 대중에게 공개하기도 했습니다(설명가능성 강화).

나아가 2023년 3월 카카오 공동체[02] 기술윤리 위원회(Tech for Good Committee)는 기존 알고리즘 윤리헌장에 대한 전면 개정에 합의하였고, 카카오와 주요 계열사의 AI 윤리 원칙을 공식적으로 **카카오 공동체의 책임 있는 AI를 위한 가이드라인**이라 확정하였습니다. 위원회 차원에서 합의한 카카오 공동체의 책임 있는 AI를 위한 가이드라인은 **카카오를 비롯한 주요 계열사가 함께 AI 윤리의 개념과 원칙에 대해 숙의하고 최종 합의**했다는 차별점을 가집니다.

거버넌스 체계의 구축

2022년 7월 **카카오 공동체 기술윤리 위원회**의 출범은 카카오의 AI 윤리 고도화를 위한 거버넌스 체계 구축의 시발점이 되었습니다. 해당 위원회는 총 8개 주요 계열사의 최고기술책임자(Chief Technical Officer, CTO)가 위원으로 참석하는 월간 회의 진행을 통해 AI 윤리와 관련된 원칙을 수립하고 자가점검 도구를 개발하는 데 기여하였으며, 기술 투명성 강화 방안에 대한 논의 또한 활발하게 이어왔습니다.

2022년 7월 위원회의 출범과 함께 **카카오 인권과기술윤리팀**이 신설되었습니다. 카카오의 인권과기술윤리팀은 AI 윤리를 포함 기술 윤리[03] 전반에 대해 구체적이고 실질적인 고민을 해왔습니다. 특히, 기술이 초래할 수 있는 프라이버시, 포용성, 투명성 문제 등 다양한 윤리적 안건들은 인권[04] 리스크를 내포하기 마련입니다. 이러한 맥락에서 카카오의 인권과기술윤리팀은 기술이 초래할 수 있는 다양한 비재무적 리스크를 인권 기반 접근(Human Rights-Based Approach, HRBA)[05] 을 통해 최소화하는데 주안점을 두고 활동하였습니다.

2023년 4월 신설된 **카카오 AI정책지원TF**는 카카오의 AI 서비스 정책 및 방향성 수립을 지원해왔으며, 2024년에 들어 카카오 AI Safety로 발전하였습니다. 기존 태스크 포스(Task Force, TF)는 기획·개발 과정에 있는 AI 서비스를 개인정보, 법무, 대외, 정책, 인권과 기술 윤리 등 다양한 관점에서 검토하여, 이를 바탕으로 전사 공통의 정책적 이슈를 발굴하는 동시에 서비스 단위에서의 실질적 이행 원칙을 마련하는데 기여하였습니다.

2024년 4월 기존 공동체 기술윤리 위원회가 CA협의체 ESG위원회 산하의 **그룹 기술윤리 소위원회**로 재편되었습니다. 또한, 기존 카카오 소속 인권과기술윤리팀은 그룹 차원의 기술윤리 정책 지원을 전담하는 카카오 CA협의체 소속 **그룹기술윤리팀**으로 발전하였습니다. 2022년 7월부터 2024년 3월까지 운영된 공동체 기술윤리 위원회가 계열사별 기술윤리를 점검하고, 발전 방안을 모색하는 컨센서스 기구인 데 반해, 그룹 기술윤리 소위원회는 리스크를 선제적으로 점검하고, 통합적으로 관리·대응하는 콘트롤타워 역할을 합니다. 기술윤리 거버넌스의 강화는 높아진 기술 기업의 사회적 책임에 걸맞은 체계와 정책을 갖추기 위한 의지가 반영된 결과입니다.

자가점검 도구의 개발

2023년 3월 카카오 공동체 기술윤리위원회가 합의한 카카오 공동체의 책임 있는 AI를 위한 가이드라인을 바탕으로 카카오 인권과기술윤리팀은 카카오와 주요 계열사가 참고할 수 있는 **AI 윤리 자가점검 도구 개발**에 매진하였습니다.

초기 도구는 **AI 윤리 체크리스트**라는 가칭으로 2023년 6월 AI 관련 사내 지식 공유 채널 중 하나인 AI 해커톤을 통해 처음 전사 공개되었습니다(간략 양식 및 통합 양식). 이후 도구의 점진적 발전을 위해 기술기획, 응용분석, 개인정보, 광고추천, 스마트채팅, 테크포털 등 다양한 부서 담당자를 대상으로 **포커스 그룹 인터뷰**(Focus Group Interview, FGI)를 진행하였으며, 이를 기반으로 도구의 발전 방향을 설정하였습니다.

카카오 공동체 기술윤리 위원회의 합의 하, 2023년 9월 기술 개발 및 서비스 직군 담당자를 포함한 **별도 카카오 AI 윤리 체크리스트 검토 TF**를 발족하였으며, 카카오브레인도 파일럿 실시를 위해 참관하였습니다.

이 설명서를 통해 소개되는 **책임 있는 AI를 위한 생애주기[06]별 자가점검 도구**는 설계 단계부터 지속적인 개선을 목표로 하였습니다. 총 63개 문항으로 구성된 해당 도구는 크게 두가지 특성을 띕니다. 우선, 문항들이 AI 시스템의 개발 단계(생애주기)에 맞춰 배치되어 있고, 카카오와 주요 계열사가 합의한 9가지 AI 윤리 원칙(카카오 공동체의 책임 있는 AI를 위한 가이드라인)과 매칭되어 있습니다. 다시 말해서, AI 시스템 관련 실무자는 도구의 각 문항 점검을 통해 AI 개발 단계에 맞는 리스크 점검을 할 수 있고, 관련 AI 윤리 원칙 이행 여부까지 확인할 수 있는 것입니다.

2024년 4월 신설된 그룹 기술윤리 소위원회는 책임 있는 AI를 위한 생애주기별 자가점검 도구를 바탕으로 **안전한 AI를 위한 핵심 체크리스트**를 개발하였습니다. 해당 체크리스트는 AI 기술과 서비스 개발 단계를 4단계(▲계획 및 설계 ▲데이터 수집 및 처리 ▲AI 모델 개발/기획 및 구현 ▲운영 및 모니터링)로 구분하고 있으며, 사회 윤리와 서비스 품질 검토에 필요한 7가지 항목으로 구성되어 있습니다. 서비스 출시 전, 카카오와 주요 계열사는 기술리더 혹은 AI리더가 점검한 체크리스트를 법무 검토 및 CEO 서명을 받아 그룹 기술윤리 소위원회에 제출해야 합니다. 책임 있는 AI를 위한 생애주기별 자가점검 도구 설명서는 사내 공개가 된 2023년도 말부터 각 계열사가 자체적으로 AI 관련 안전 점검 절차를 수립하거나 도구를 개발할 때 참고할 수 있는 자료로 활용되고 있습니다.

1 여는 글: 카카오의 AI 윤리 거버넌스 고도화를 향한 과정

안전한 AI를 위한 핵심 체크리스트에 이르기까지 도구를 다듬어가는 과정에서 스태프 직군, 서비스 직군, 테크 직군 등 카카오 내 다양한 직군 간 활발한 논의가 진행되었으며, 해당 설명서에 이러한 다수 이해관계자 간 AI 윤리 토의 과정을 담았습니다. **다양한 이해관계자 간 소통**은 공통의 목적을 민주적이고 통합적으로 달성하기 위한 기반이 됩니다. 특히 기술을 대변하는 AI와 사회과학·인문을 대변하는 윤리라는 다소 이질적으로 보일 수 있는 두 개념을 조화롭게 검토하기 위해서는 다양한 지식과 경험을 겸비한 직군 간 교류가 필수적인 부분이었다고 할 수 있습니다.

의의 및 개선 방향

이 설명서는 **카카오의 AI 윤리 원칙 수립 및 실천 도구의 개발 과정**을 담고 있습니다. AI 윤리 점검 도구의 제작 과정을 공유하는 것은 카카오가 지속적으로 도구의 발전 및 활용의 계기를 마련하고, 이를 통해 안전하고 신뢰할 수 있는 AI 서비스 개발을 할 수 있게 하는 데 의의가 있습니다. 덧붙여, AI 윤리 거버넌스 고도화의 경험 공유를 통해 기술 기업으로서 카카오의 사회적 책임 실천의 사례를 알리고, AI 윤리를 내재화하려는 타 조직에게 도움이 될 수 있습니다.

AI 윤리 점검 도구는 기술의 발전과 사회의 변화에 맞춰 변화해야 합니다. 무엇보다 AI 윤리를 포함한 **전사 차원에서의 AI 사업 방향성 및 거버넌스 체계 구체화**가 필요하고, **자가점검 도구의 실제 서비스 적용**을 통한 도구의 **세부 문항 개선**도 이루어져야 합니다. 덧붙여, **도구의 사용 시점, 점검 이행 절차** 및 **담당자 지정** 등 많은 과제가 전사 차원에서 다루어져야 합니다. 이러한 맥락에서 **AI 윤리 자가점검 도구는** 지속적인 개선 작업을 거칠 예정입니다.

주요 개념 정리

01 AI 윤리

- 기술 윤리에 포함되는 개념
- 추천 알고리즘, 생성형 AI 등 AI 기술을 개발·활용하면서 대두되는 윤리적 안건(데이터 프라이버시, 알고리즘 편향성, 차별 등)에 초점을 맞춘 도덕적 관념

02 카카오 공동체

- 카카오와 주요 계열사(카카오게임즈, 카카오모빌리티, 카카오뱅크, 카카오브레인, 카카오엔터프라이즈, 카카오엔터테인먼트, 카카오페이)를 통칭하는 개념
- 2024년 기준, 기존 카카오 공동체를 카카오그룹으로 칭하고 있으며, 카카오브레인이 카카오에 합병, 카카오헬스케어가 새로 참여

03 기술 윤리

- 다양하고 복잡해지는 기술의 변화 속에서 새롭게 대두되는 윤리적 문제를 인식하고 선제적으로 대응하는 데 필요한 도덕적 관념

04 인권

- 세계인권선언 제1조 및 제2조: 모든 인간은 태어날 때부터 자유로우며 그 존엄과 권리에 있어 동등하다. 모든 사람은 인종, 피부색, 성, 언어, 종교, 정치적 또는 기타의 견해, 민족적 또는 사회적 출신, 재산, 출생 또는기타의 신분과 같은 어떠한 종류의 차별이 없이, 이 선언에 규정된 모든 권리와 자유를 향유할 자격이 있다.
- 대한민국 헌법 제10조: 모든 국민은 인간으로서의 존엄과 가치를 가지며, 행복을 추구할 권리를 가진다. 국가는 개인이 가지는 불가침의 기본적 인권을 확인하고 이를 보장할 의무를 진다.
- 국가인권위원회법 제1장 제2조: 인권이란 대한민국헌법 및 법률에서 보장하거나 대한민국이 가입·비준한 국제인권조약 및 국제관습법에서 인정하는 인간으로서의 존엄과 가치 및 자유와 권리를 말한다.

05 인권 기반 접근(Human Rights-Based Approach, HRBA)

- 국제인권기준과 원칙을 개발 계획, 정책 결정, 시행과 평가 전 과정에 통합하는 개념적 틀

06 생애주기

- 소프트웨어가 기획, 개발 및 배포되는 일련의 공정을 체계화한 절차
- 해당 설명서에서 언급하는 AI 시스템의 생애주기는 다음과 같은 단계로 이루어져 있음
  ① 계획 및 설계
  ② 데이터 수집 및 처리
  ③ AI 모델 개발
  ④ 시스템 기획 및 구현
  ⑤ 운영 및 모니터링

1

카카오가 추구하는 책임 있는 AI는 AI 서비스 초기부터 다양한 직군의 실무자가 자가점검을 통해 AI 사회적 리스크를 최소화하는 것을 목표로 합니다.

2

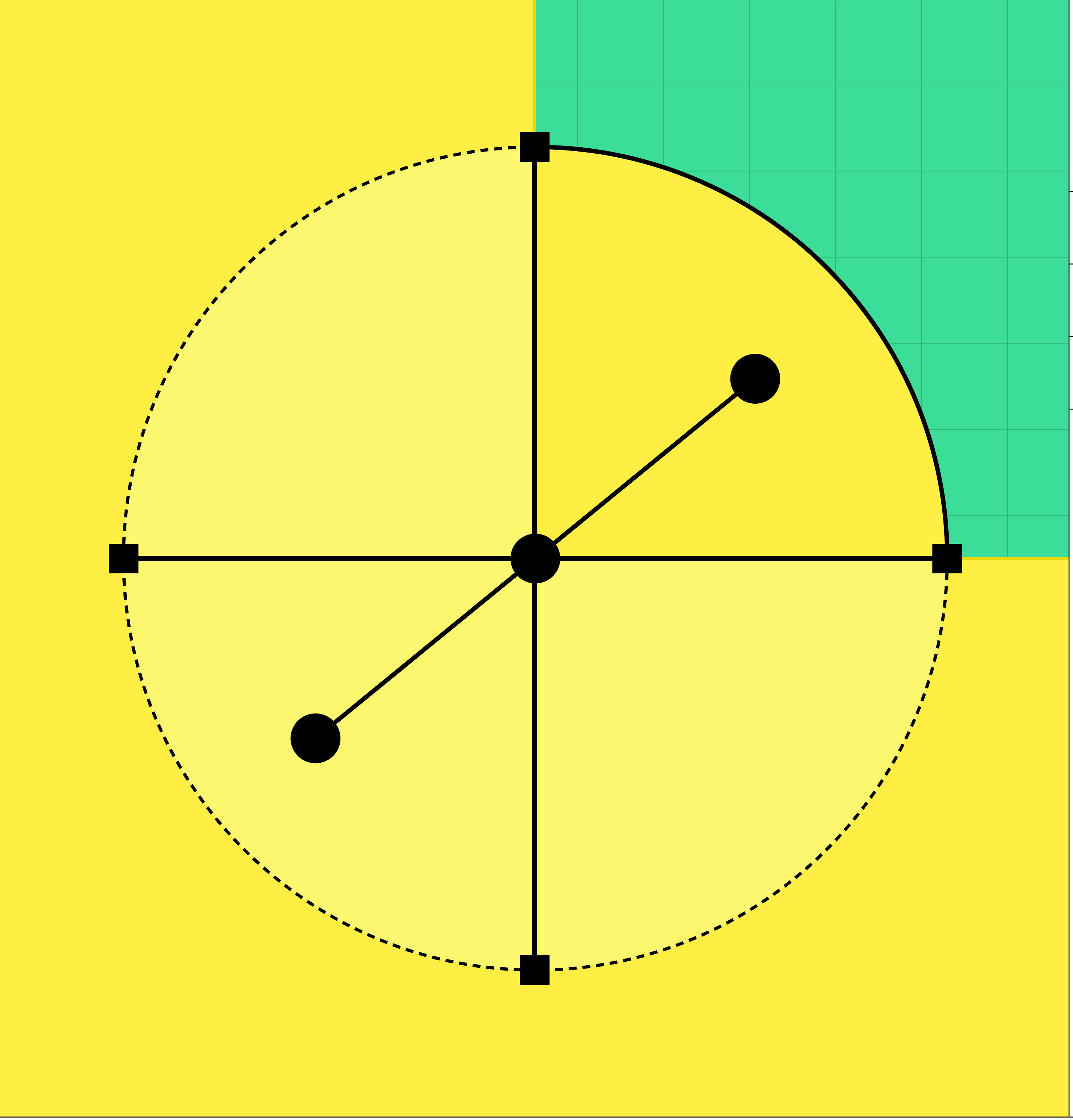
카카오와 카카오 주요 계열사가 합의한 AI 윤리 원칙을 바탕으로 원칙의 실천을 도울 수 있는 자가점검 도구를 개발하였습니다.

3

향후 다양한 서비스 적용을 통해 도구의 효과성을 보완하고자 하며, AI 거버넌스 전반을 고도화하기 위해 노력하고자 합니다.

앞서 설명했듯이 2024년 4월 신설된 그룹 기술윤리 소위원회는
책임 있는 AI를 위한 생애주기별 자가점검 도구를 바탕으로
안전한 AI를 위한 핵심 체크리스트를 개발하였습니다.

**해당 페이지 이후의 내용은 2023년도 말까지 개발된
책임있는 AI를 위한 생애주기별 자가점검 도구의 개발 과정 및 의의를
기술해둔 것입니다.**



카카오의 AI 윤리 거버넌스 고도화 현황

2

2 카카오의 AI 윤리 거버넌스 고도화 현황

카카오는 다년간 카카오만의 책임 있는 AI 실천 방안을 고민해 왔습니다. 정책 수립, 담당 조직 및 협의체 구성, 도구의 개발 등 오늘도 카카오는 AI 현장에서 AI 윤리를 내재화하기 위해 노력하고 있습니다. 다음은 2018년부터 2023년 말까지 이어진 카카오의 AI 윤리 거버넌스 고도화 현황을 가시화한 것입니다.

카카오의 AI 윤리 거버넌스 고도화 현황: 타임라인

2018년 1월	2022년 7월	2023년 3월	2023년 4월	2023년 11월
알고리즘 윤리헌장 공개	인권과기술윤리팀 설립 카카오 공동체 기술윤리 위원회 설립	카카오 공동체의 책임 있는 AI를 위한 가이드라인 수립	카카오 AI정책지원TF 설립	책임 있는 AI를 위한 생애주기별 자가점검 도구 개발

카카오 알고리즘 윤리헌장 공개

2018년 1월 **카카오 알고리즘 윤리헌장**을 공개한 후, 카카오는 포용성(2019년), 아동과 청소년 보호(2022년) 및 프라이버시(2022년) 관련 조항을 추가하는 등 끊임없이 AI 윤리 고도화를 향해 전진하고 있습니다. 예컨대, 증오발언 근절 원칙을 이모티콘 제작 가이드에 적용하여 AI 시스템이 초래할 수 있는 차별에 대한 경계를 실천하고, 다음 뉴스 배열 설명서를 외부 이해관계자에게 공개하여 AI 서비스의 투명성을 강화하기도 했습니다.

카카오 알고리즘 윤리헌장 (2018년 1월)

원칙		상세 내용
1	**카카오 알고리즘의 기본원칙**	카카오는 알고리즘과 관련된 모든 노력을 우리 사회 윤리 안에서 다하며, 이를 통해 인류의 편익과 행복을 추구한다.
2	**차별에 대한 경계**	알고리즘 결과에서 의도적인 사회적 차별이 일어나지 않도록 경계한다.
3	**학습 데이터 운영**	알고리즘에 입력되는 학습 데이터를 사회 윤리에 근거하여 수집·분석·활용한다.
4	**알고리즘 독립성**	알고리즘이 누군가에 의해 자의적으로 훼손되거나 영향받는 일이 없도록 엄정하게 관리한다.
5	**알고리즘에 대한 설명**	이용자와의 신뢰 관계를 위해 기업 경쟁력을 훼손하지 않는 범위 내에서 알고리즘에 대해 성실하게 설명한다.
6	**기술의 포용성**	알고리즘 기반의 기술과 서비스가 우리 사회 전반을 포용할 수 있도록 노력한다.
7	**아동과 청소년에 대한 보호**	카카오는 아동과 청소년이 부적절한 정보와 위험에 노출되지 않도록 알고리즘 개발 및 서비스 디자인 단계부터 주의한다.
8	**프라이버시 보호**	알고리즘을 활용한 서비스 및 기술의 설계와 운영 등의 전 과정에서 이용자의 프라이버시 보호에 소홀함이 없도록 노력을 다한다.

2.1 정책 개선

카카오 공동체의 책임 있는 AI를 위한 가이드라인 수립

2023년 3월 카카오 공동체 기술윤리 위원회는 기존 알고리즘 윤리헌장을 바탕으로 생성형 AI의 등장 등 현 시류에 맞게 AI 윤리 원칙을 개선하였으며, 그 공식 명칭을 **카카오 공동체의 책임 있는 AI를 위한 가이드라인**이라 확정하였습니다.

카카오 공동체의 책임 있는 AI를 위한 가이드라인 (2023년 3월)

원칙		상세 내용
1	사회 윤리	카카오 공동체(이하 '공동체')는 인공지능(이하 'AI')과 관련된 모든 노력을 우리 사회 윤리 안에서 다하며, 이를 통해 인류의 편익과 행복을 추구한다.
2	비차별과 비편향	공동체는 AI가 성별, 인종, 출신, 종교 등을 이유로 차별을 초래하거나 이에 근거한 편향된 의사결정을 하지 않도록 경계한다.
3	포용성	공동체는 소외 없는 AI 기술 제공을 디지털 기업의 중요한 책임으로 인식하고, 연령과 장애 여부 등과 상관없이 모든 이용자가 공동체 AI 기술을 동등하게 이용할 수 있는 환경을 추구한다. 공동체는 AI 기술과 서비스가 우리 사회 전반을 포용할 수 있도록 노력한다.
4	투명성	공동체는 서비스 내 AI 기술의 활용 목적, 활용 내용 및 위험 요소에 대해서 이용자에게 지속적으로 알린다. 이를 위해 구성 요소들을 기업 경쟁력을 훼손하지 않는 범위 내에서 성실하게 설명한다.
5	보안과 안전, 독립성	공동체는 AI 시스템과 데이터를 외부 공격 및 악의적인 사용과 같은 위험에서 방어할 수 있도록 법적 요구사항을 준수하는 보안 체계를 구축한다. AI 시스템의 비정상적 작동과 예기치 못한 상황에 대한 안전조치와 대응 체계를 마련해 안정적인 서비스가 제공되도록 노력한다. AI가 외부적 혹은 내부적 요인으로 훼손되거나 영향 받는 일이 없도록 엄정하게 관리한다.
6	인권	공동체는 AI 기술 개발과 활용 전 과정에서 인권 존중의 책임 이행을 목표로 한다. AI 기술로 인해 이해관계자의 권리가 보장받지 못하는 상황에 대해 예방 및 대응 조치를 마련할 수 있도록 노력한다.
7	프라이버시	공동체는 AI 기술 개발과 활용 전 과정에서 이용자의 프라이버시를 보호한다. 개인정보 오용을 최소화하기 위해 관련 법규에 따라 개인정보 보호 절차를 마련하고, 예방 및 대응조치를 취한다.
8	이용자 보호	공동체는 AI 기술에 각종 법과 규범이 요구하는 이용자 보호 정책을 효율적으로 반영한다. 이용자의 권익이 침해되지 않도록 안전한 예방 절차를 수립하기 위해 노력하고 이후에도 관련 절차가 제대로 작동하는지 지속적으로 확인한다.
9	역기능에 대한 경계	공동체는 AI 기술의 불완전성을 인정하고, 설계 의도와 다른 사회적 결과가 발생할 가능성에 대비한다. 필요시 관련 전문가와 협력을 통해 해결책을 모색한다.

2.2 거버넌스 체계 구축

카카오 공동체 기술윤리 위원회 설립

2022년 7월 카카오 공동체 기술윤리 위원회의 출범은 카카오의 AI 윤리 고도화를 위한 거버넌스 체계 구축의 시발점이 되었습니다. 카카오, 카카오게임즈, 카카오페이, 카카오엔터테인먼트, 카카오모빌리티, 카카오엔터프라이즈, 카카오브레인, 카카오뱅크까지 총 8개 주요 계열사의 CTO가 위원으로 참석하는 월간 회의를 통해 카카오와 주요 계열사는 AI 윤리 관련 원칙 수립, 자가점검 도구 개발, 투명성 강화 등에 대해 활발하게 논의해왔습니다.

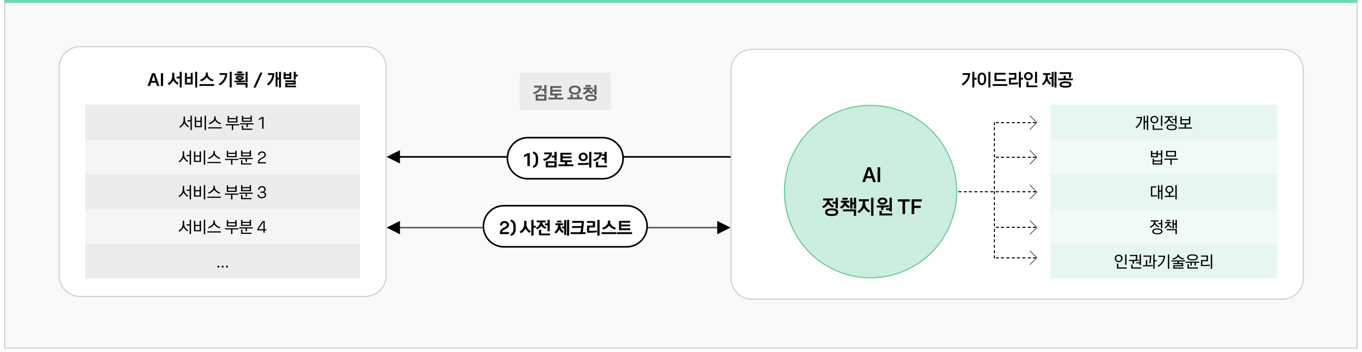
카카오 인권과기술윤리팀 설립

2022년 7월 위원회의 출범과 함께 **카카오 인권과기술윤리팀**이 신설되었습니다. AI 윤리를 포함하여 기술 윤리 전반에 대해 실질적으로 고민하는 전담 조직 및 담당자가 지정된 것은 기술 윤리 고도화에 대한 카카오의 실천 의지를 보여줍니다.

카카오 AI정책지원TF 설립

2023년 4월에 신설된 카카오 **AI정책지원TF**는 카카오의 AI 서비스 정책 및 방향성 수립을 지원하는 역할을 했습니다. 해당 TF를 통해 카카오는 기획·개발 과정에 있는 AI 서비스를 개인정보, 법무, 대외, 정책, 인권과 기술 윤리 등 다양한 관점에서 검토하였습니다.

카카오 AI정책지원TF(2023년)의 업무 체계



2.3 자가점검 도구 개발

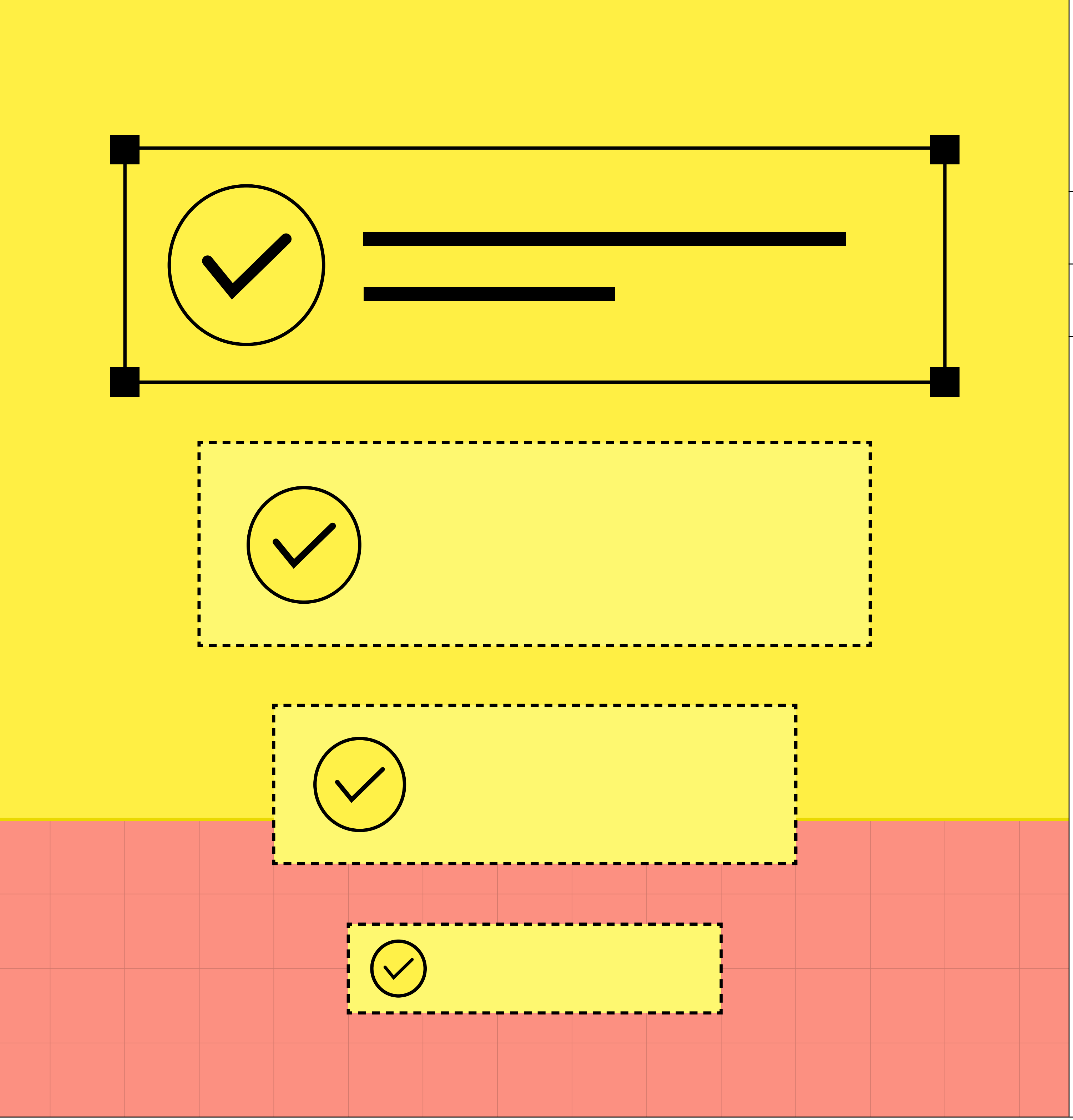
책임 있는 AI를 위한 생애주기별 자가점검 도구 개발

2023년 3월 카카오 공동체 기술윤리위원회가 합의한 **카카오 공동체의 책임 있는 AI를 위한 가이드라인**을 바탕으로 약 6개월 간 카카오 AI 윤리 체크리스트 검토 TF는 카카오와 주요 계열사가 참고할 수 있는 **AI 윤리 자가점검 도구** 개발에 매진하였습니다.

TF의 구성을 보면, AI 서비스와 관련된 카카오(테크 직군, 서비스 직군, 스태프 직군) 및 카카오브레인의 주요 이해관계자들로 구성되어 있습니다.

책임 있는 AI를 위한 생애주기별 자가점검 도구 개발 과정

- **AI 윤리 체크리스트: 간략 및 통합 양식 개발**(2023년 6월)
- **AI 시스템의 생애주기를 반영한 도구로 발전**(2023년 9월)
- **카카오 AI 윤리 체크리스트 검토 TF 발족**(2023년 9월)
- **TF 1차 검토 결과 반영**(2023년 9월)
- **카카오브레인 파일럿 실시**(2023년 10월)
- **TF 2차 검토 결과 반영**(2023년 10월~11월)
- **내부 회람 및 도구 확정**(2023년 11월~2024년 1월)



책임 있는 AI를 위한 생애주기별 자가점검 도구

3

3.1 도구에 대한 소개

점검 대상

책임 있는 AI를 위한 생애주기별 자가점검 도구가 적용될 대상을 '점검 대상'이라 규정하고 있으며, 이는 간단히 말해 카카오가 '개발·활용하는 AI 시스템'을 지칭합니다. 여기에서 말하는 **AI 시스템**이란, **주어진 목표에 따라 실제 환경이나 가상 환경에 영향을 미치는 예측, 추천 또는 의사결정과 같은 출력물을 생성할 수 있는 엔지니어링 또는 기계 기반 시스템**을 의미합니다(출처: 미국 국립표준기술원의 AI 리스크 관리 프레임워크).

점검 요소

AI 시스템의 생애주기란, **AI 시스템을 구성하는 데이터, 모델 등의 요소를 구현하고 운영하는 과정**을 일컫습니다. 책임 있는 AI 시스템을 확보하기 위해서는 시스템의 구상부터 배포까지 전 단계에 걸쳐서 카카오 공동체의 책임 있는 AI를 위한 가이드라인 준수 여부를 확인해야 합니다. 해당 도구는 AI 시스템의 생애주기(5단계)에 맞추어 각 단계에서 다뤄야 할 윤리 문제들을 고려하도록 하였으며, 현재 총 63개 세부 문항으로 구성되어 있습니다. 또한 각 세부 문항을 카카오 공동체의 책임 있는 AI를 위한 가이드라인 9개 항목과 매칭시킴으로써 카카오와 주요 계열사가 합의한 AI 윤리 원칙별 실천 현황을 한눈에 확인할 수 있게 하였습니다.

필요시, 점검 담당자는 각 세부 문항별 '주요 개념,' '주요 지표,' '근거' 및 '질문의 의도'를 참고할 수 있으며, 이는 문항별로 적절한 답변과 근거자료를 준비하는 데에 유용한 지침이 될 수 있습니다. 점검 담당자는 특히 주요 지표를 기준으로 '예,' '아니요,' '해당 없음' 중 답변을 선택할 수 있으며, 제시된 근거에 부합하는 해당 답변을 뒷받침할 자료를 첨부할 수 있습니다. 답변이 '예'일 경우, 정책, 활동 보고서 등 실천 현황을 증빙하는 자료를 첨부하고, '아니요'일 경우, 그 이유와 앞으로의 개선 계획을 첨부해야 합니다. '해당 없음'일 경우에는 사유를 설명하는 방식으로 답변 및 근거를 작성할 수 있습니다.

점검 주체

해당 도구를 활용하여 **점검을 시행할 주체**는 AI 시스템의 특성, 우선순위 위험 요소, 조직 내 정책 및 업무 규정 등에 대해 충분히 이해하고 있는 자입니다(추후 도구 적용 예정인 서비스별 담당자 지정 및 지속적 교육 필요). 하지만 보다 넓은 관점에서 볼 때, 점검을 시행하는 과정에서 AI 시스템을 구현하는 과정에 직·간접으로 관련되거나 영향을 주는 조직과 개인 모두가 도구에 대한 문해력을 갖는 것이 중요합니다.

3.1 도구에 대한 소개

시행 체계

시행 체계란, **점검 담당자 입장에서의 자가점검 도구 활용법**을 의미합니다.

도구 활용법

① AI 시스템 기획·개발 초기 단계에 **AI 시스템의 생애주기별 윤리 요건**에 대해 전반적으로 숙지
- [1단계] 계획 및 설계: AI 시스템 전반에 대한 이해, 우선순위 위험 식별 및 위험관리 계획 수립, 이해관계자 조사
- [2단계] 데이터 수집 및 처리: 학습 데이터 확보 과정에서 발생할 수 있는 데이터 오류 및 편향에 대한 관리 방안 확보
- [3단계] AI 모델 개발: 학습 모델의 편향적인 출력이나 공격에 대응하는 방안 수립, 학습 모델의 출력을 해석하는 방안 제공
- [4단계] 시스템 기획 및 구현: AI 시스템 개발 시 발생할 수 있는 편향이나 오류에 대한 대응책 마련, AI 서비스가 도출한 결과에 대해 사용자 친화적인 설명 제공
- [5단계] 운영 및 모니터링: AI 시스템 문제 발생 시 원인 추적을 통해 대응 방안 마련

② **1단계인 '계획 및 설계'**를 통해 AI 시스템 전반에 대한 이해, 우선순위 위험 식별 및 위험관리 계획 수립, 이해관계자 조사 등을 진행

③ **세부 문항**을 읽고, 답변 선택**(예, 아니요, 해당 없음)**. 정확한 답변을 위해 각 세부 문항별 **'주요 개념', '주요 지표', '근거', '질문의 의도'** 참고

④ 선택한 답변에 대한 **근거** 기재

⑤ 세부 문항별 답변을 통해 **카카오 공동체의 책임 있는 AI를 위한 가이드라인 준수 여부 확인**

도구 활용을 위한 AI 윤리 점검 체계 마련의 필요성

카카오는 AI 시스템 생애주기 전반에 걸쳐 테크 직군, 서비스 직군, 스태프 직군 등 다양한 직무의 이해관계자가 유기적으로 협력할 수 있는 AI 윤리 업무 체계를 구축하기 위해 노력해왔습니다. 2023년 12월 기준, **카카오 AI정책지원TF 업무 체계**(카카오의 AI 서비스 정책 및 방향성 수립을 지원하는 역할을 하며, 기획·개발 과정에 있는 AI 서비스를 개인정보, 법무, 대외, 정책, 인권과 기술 윤리 등 다양한 관점에서 검토) 및 **AI 윤리 점검 체계**(인권과기술윤리팀, 카카오 AI정책지원TF, 카카오 공동체 기술윤리 위원회, ESG 총괄 위원회 등 전사 차원에서 AI 윤리 업무 검토)를 중심으로 AI 윤리 업무 체계가 고도화되었습니다. 이를 바탕으로 카카오는 AI 윤리 점검 도구의 활용이 실제 구현될 수 있도록 노력하고 있습니다. **부서 간 협업을 활성화할 수 있는 AI 업무 체계**는 원활한 AI 윤리 점검의 전제 조건이라 할 수 있습니다. 카카오는 이번 설명서에서 소개하는 책임 있는 AI를 위한 생애주기별 자가점검 도구를 통해 **AI 시스템 기획·개발 초기 단계부터** 다양한 직군의 실무자들이 함께 AI 윤리에 대해 함께 고민하기는 토대를 만들고자 했습니다.

3.2 도구의 개발 과정

도구의 개발 과정(2023년)

3월
카카오 공동체의 책임 있는 AI를 위한 가이드라인 수립
카카오 공동체
기술윤리 위원회

6월
AI 윤리 체크리스트 개발
간략/통합 양식

9~11월
AI 시스템의 생애주기를 반영한 도구로 개선

9월
카카오 AI 윤리 체크리스트 검토 TF 발족
- 카카오 테크/서비스 직군
- 카카오브레인

카카오브레인 파일럿 실시

10월
TF 1차 검토에 대응
- 문항 상세화
- 용어집 제작
- 문항 간명화(112→99)

11월
TF 2차 검토에 대응
- 문항 수 최소화 (99→63)
- 향후 방향성 도출

12월
도구의 확정 및 설명서 초안 공유
책임 있는 AI를 위한
생애주기별 자가점검 도구

3.2 도구의 개발 과정

카카오 공동체의 책임 있는 AI를 위한 가이드라인 수립

최근 생성형 AI 기술의 급격한 발전은 대중의 관심과 기술의 상용화로 이어지는 추세입니다. 다른 한편, AI 기술의 부작용과 적절한 대처 결여에 대한 우려의 목소리 또한 높아지고 있으며, 규제 당국은 기술 기업에 이전보다 높은 수준의 책임감과 구체적인 이행 원칙을 요구하고 있습니다.

카카오는 2018년 **알고리즘 윤리헌장을 발표**하여 카카오만의 원칙과 철학에 기반한 알고리즘 개발 및 운영에 관한 의지를 표명한 바 있습니다. 이후 2023년 카카오와 주요 계열사는 카카오 공동체 기술윤리 위원회 의결 과정을 통하여 **카카오 공동체의 책임 있는 AI를 위한 가이드라인** 즉, 개선된 AI 윤리 원칙을 확립하였습니다. 총 9개 항목으로 이루어진 해당 가이드라인은 카카오뿐만 아니라 주요 계열사가 핵심 가치를 공유하고, 그 가치를 실제 서비스에 반영하겠다는 의지를 선언한다는 점에서 차별성을 가집니다.

카카오와 주요 계열사는 가이드라인 수립에 착수하기 전부터 **AI 윤리 원칙의 실천을 극대화**할 방안을 모색하는 것에 방점을 두었습니다. 카카오 공동체 기술윤리 위원회는 각 카카오 계열사의 CTO가 모인 자리인 만큼 기술 개발의 현장에서 포착된 실제 AI 윤리 사안을 공유할 수 있는 자리가 되었고, 여기서 공유된 AI 윤리에 대한 다양한 기술적 관점을 참작하여 가이드라인이 수립되었습니다. 더불어 카카오와 주요 공동체는 해당 가이드라인의 아홉 개 원칙을 기본 골자로 삼아 각 계열사에 적합한 방식으로 AI 윤리를 내재화 및 고도화하기 위한 자가점검 도구를 제작하기로 하였으며, 이는 이번 설명서에서 소개되는 책임 있는 AI를 위한 생애주기별 자가점검 도구의 초석이 되었습니다.

3.2 도구의 개발 과정

AI 윤리 체크리스트 개발

2023년 6월, 카카오 인권과기술윤리팀은 **AI 윤리 체크리스트**를 준비하여 카카오 공동체 기술윤리 위원회와 전사 AI 지식 공유 채널에 소개하였습니다. 이는 해당 설명서에서 가장 최근 성과물로 소개되는 책임 있는 AI를 위한 생애주기별 자가점검 도구의 초안이라고 할 수 있습니다.

그 과정에서 AI에 대한 '기술적 고려 요소' 및 '인권 리스크 점검 요소'를 아우르는 도구를 만들기 위해 '2023 인공지능 윤리 기준 실천을 위한 자율점검표(2023. 과학기술정보통신부, 정보통신정책연구원)'와 '인공지능 인권영향평가 도입 방안연구(2022. 국가인권위원회)'을 주로 참고하였습니다.

결과적으로 두 가지 양식의 산출물을 낼 수 있었습니다. **통합 양식**은 인권 기반 접근(Human Rights-Based Approach, HRBA)을 통해 AI 윤리 문제를 해결하고자 고안한 도구로 인권영향평가의 틀(단계 1. 계획 및 준비; 단계 2. 분석 및 평가; 단계 3. 개선 및 구제; 단계 4. 모니터링 및 공시)을 최대한 반영하였습니다. 하지만 양이 방대하였고, 기술 분야 실무자가 보았을 때 생소한 개념들도 있었습니다.

통합 양식의 방대한 양과 개발자에게 생소한 용어를 최소화하기 위해 무엇보다 체크리스트 사용자가 카카오의 AI 윤리 원칙(카카오 공동체의 책임 있는 AI를 위한 가이드라인)에 따른 점검 사항을 직관적으로 쉽고 빠르게 숙지할 수 있도록 통합 양식의 일부를 발췌하여 **간략 양식**을 준비하게 되었습니다.

3.2 도구의 개발 과정

AI 시스템의 생애주기를 반영한 도구로 발전

카카오 인권과기술윤리팀은 기존 개발된 **자가점검 도구 초안(AI 윤리 체크리스트: 간략 양식 및 통합 양식)이 갖는 한계를 보완**하기 위해 고민해 왔고, **2023년 9월 한층 발전된 도구를 준비**할 수 있었습니다.

자가점검 도구 초안의 한계 분석

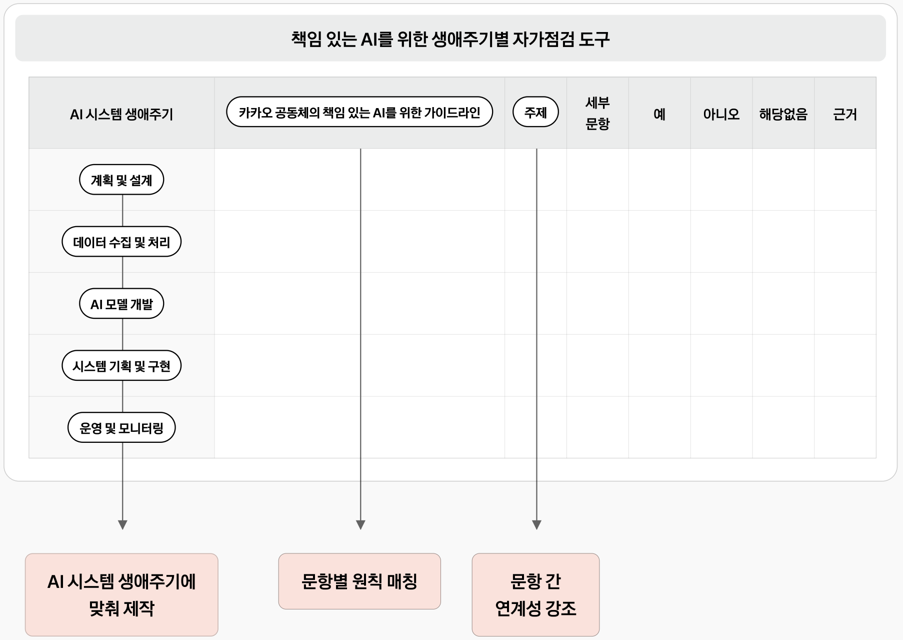
- **국가인권위원회가 제공하는 인공지능 인권영향평가의 프레임**(단계 1.계획 및 준비; 단계 2.분석 및 평가; 단계 3.개선 및 구제; 단계 4.모니터링 및 공시)을 바탕으로 개발되어, 해당 프레임에 익숙하지 않은 실무자가 사용하기에 **문항 수가 많고 실용적이지 않다**고 느껴질 수 있음
- **간략과 통합 두 가지 양식**을 동시에 제시할 경우, 도구를 활용하여 점검을 시행할 주체는 언제 어떤 도구를 활용해야 하는지 **혼란스러울 수 있음**
- 인권, AI 윤리 등 생소한 개념이 등장하는 **세부 문항의 의미 그리고 각 문항 간 연계성에 대한 설명이 보완**될 필요성이 있음
- 도구의 활용을 통해 카카오와 주요 계열사가 합의한 AI 윤리 원칙인 **카카오 공동체의 책임 있는 AI를 위한 가이드라인 준수 여부를 확인**하는 것이 쉽지 않음

자가점검 도구 초안과 비교하여 개선된 부분

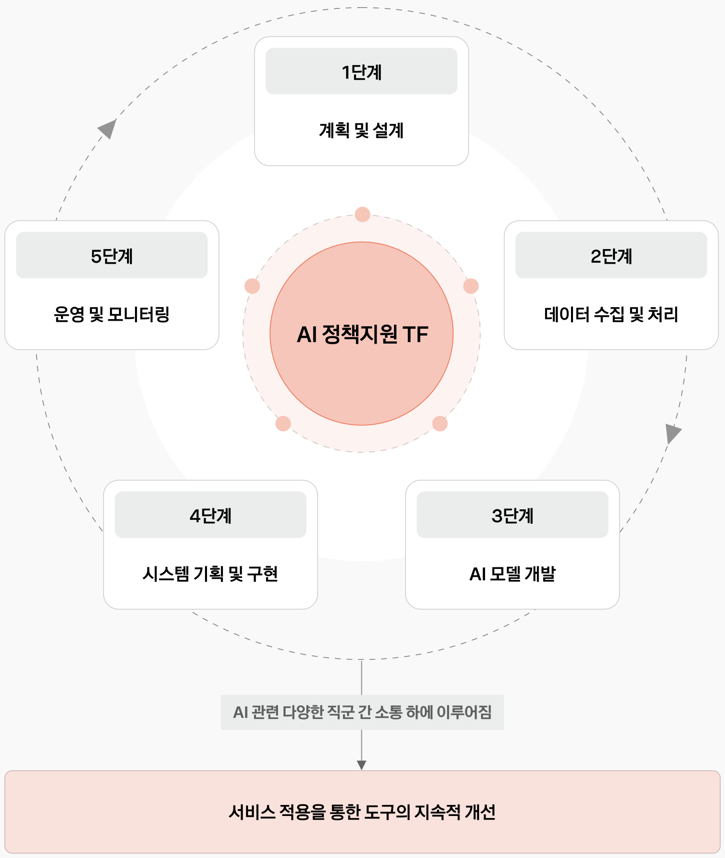
- **인권영향평가의 내용을 반영하되 도구의 구성을 'AI 시스템 생애주기'에 맞춰 제작**
  - 실무자가 AI 시스템의 기획·개발 초기 단계부터 AI 시스템 생애주기에 따라 AI 윤리 점검 요소를 확인할 수 있도록 도구를 개선
  - 이때, 추가로 2022 신뢰할 수 있는 인공지능 개발 안내서(안)(2022. 과학기술정보통신부, 한국정보통신기술협회)를 참고
- **도구의 세부 문항을 분류하는 '주제'란 추가**
  - 세부 문항의 의미 및 문항 간 연계성을 파악할 수 있도록 주제란을 추가
- **세부 문항별로 '카카오 공동체의 책임 있는 AI를 위한 가이드라인' 매칭**
  - 문항의 답변을 카카오 공동체의 책임 있는 AI를 위한 가이드라인의 항목과 매칭할 수 있게 함으로써 어떤 AI 윤리 원칙을 준수하고 있는지 확인할 수 있도록 함

3.2 도구의 개발 과정

AI 시스템의 생애주기를 반영한 도구: 프레임워크



도구의 활용과 개선



3.2 도구의 개발 과정

카카오 AI 윤리 체크리스트 검토 TF 발족

2023년 9월, **AI 시스템 생애주기에 맞춰 개선한 도구**를 보다 다양한 내부 이해관계자에게 검토받고자 **카카오 AI 윤리 체크리스트 검토 TF를 구성**하게 되었습니다. 서로 다른 직군의 카카오 그리고 카카오브레인 실무자가 도구를 검토함으로써 다양한 관점에서 실무적 보완점을 발견할 수 있었으며, 이는 AI 기술 개발의 현장에서 효과적으로 활용될 수 있는 도구를 개발하는 데 큰 역할을 하였습니다. 특히, **카카오브레인이 인턴십 프로그램을 통해 기획한 두 개의 서비스를 대상으로 도구를 시범 적용**해 얻은 결과는 도구의 유용성과 실효성을 검증하는 데에 박차를 가했습니다.

TF의 구성

AI 서비스와 관련된 카카오(테크 직군, 서비스 직군, 스태프 직군) 및 카카오브레인의 주요 이해관계자들이 참석자 또는 참관인의 자격으로 TF에 참여하였습니다.

TF의 도구 검토: 진행 일정

2023년 9월부터 11월까지 이어진 3개월간의 **카카오 AI 윤리 체크리스트 검토 TF의 업무 진행 일정**은 다음과 같습니다.

[1] **1차 검토** (9월 8일~9월 22일): **AI 시스템 생애주기에 맞춰 개선한 도구** 1차 검토

[2] **카카오브레인 파일럿** (10월 4일~10월 31일): 1차 검토를 토대로 개선한 도구를 **인턴십 프로그램에서 개발한 2개 서비스에 시범 적용**

[3] **2차 검토** (10월 31일~11월 9일): **카카오브레인 파일럿을 통해 도출된 결과**를 토대로 도구 2차 검토(최종 검토)

3.2 도구의 개발 과정

TF 1차 검토 결과 반영

2023년 9월 8일부터 9월 22일까지 **AI 시스템 생애주기에 맞춰 정리한 도구에 대한 TF의 1차 검토**가 이루어졌으며, 그 결과는 다음과 같습니다.

도구를 실제 활용하거나 직·간접적으로 영향을 받게 될 이해관계자의 의견을 충분히 이해하고 반영해야만 도구 사용자에게 부담이 아닌 도움이 되는 도구를 만들 수 있고, 비로소 AI 윤리를 기술 개발 현장에서 실천할 수 있습니다. 따라서, **TF의 1차 검토 의견을 종합 및 참작**하여 다음과 같이 도구 개선 방향을 설정하고 수정하였습니다.

도구 개선 방향의 설정과 수정

검토 전	검토 후
불명확한 용어의 의미 '안전성' 혹은 '위험 및/또는 위험 요소' 등 의미에 있어서 해석의 범위가 넓은 용어는 그 의미가 정의될 필요가 있음	**주요 개념** 문항에 언급된 개념이나 용어 중 명확하지 않거나 중요한 것에 대한 설명, 정의 및 예시 제공
지표의 부재 '예,' '아니요,' '해당 없음' 중 한 답변을 선택하거나 '충분히 분석'되었는가 등 정도를 묻는 질문에 답하기 위해 기준이 명시될 필요가 있음	**주요 지표** 문항에 대해 '예,' '아니요,' '해당 없음' 중 하나의 답변을 선정하도록 참고할 수 있는 기준 제시
문항의 의도나 맥락 파악의 어려움 메타데이터 생성 여부를 묻는 등 구체적인 기술적 내용을 확인하는 문항의 의도나 맥락이 파악하기 어려워 이에 대한 보완 설명이 필요함	**근거** 문항에 대한 답변을 뒷받침할 수 있는 구체적인 증거의 예시 제공
예시 및 참고문헌에 대한 안내 부재 도구의 답변에 대한 근거자료로 사용될 수 있는 예시 및 참고문헌에 대한 안내가 필요	**질문의 의도** 문항이 삽입된 이유나 배경, 문맥에 대한 상세한 설명 제공

문항 상세화

기존 문항에 '주요 개념,' '주요 지표,' '근거,' '질문의 의도'를 추가함으로써 문항 답변에 필요한 정보를 상세하게 전달

카카오브레인 파일럿 실시

카카오브레인은 TF 1차 검토 의견을 바탕으로 개선된 도구를 자사 인턴십 프로그램을 통해 기획한 두 개의 **AI 기반 서비스에 시범 적용**하였습니다.

AI 윤리를 확보하기 위한 전사 차원의 준비 필요

- AI 윤리를 확보하기 위한 전사 차원의 준비가 필요함
- 예를 들어, 자가점검 도구의 일부 문항들은 AI 윤리와 관련된 위험 요소를 사전에 식별하고 발생 가능한 위험에 대처하기 위한 체계가 마련되었는지를 확인함. 단, 자가점검 시행 후에 이러한 부분이 미흡하다는 것을 발견하더라도 협업 체계가 부재하거나 AI 윤리 점검자가 미리 지정되어 있지 않으면 직접 그리고 당장 행동할 수 없는 경우가 대다수임
- 따라서, 위험 관리 체계를 마련하기 위한 (1) 구체적인 방안을 제시하거나 (2) 조직 차원에서 문제 해결에 직접 착수함이 필요하며 (3) 필요시 자문할 수 있는 전문인력을 제안해야 하는 등 많은 준비가 사전에 전사 차원에서 이루어져야 함

서비스별 우선순위 설정 및 도구 최적화 필요

- AI 시스템은 다양하기 때문에 (1) 도구를 어떤 서비스에 우선 적용할지, (2) 서비스별로 도구를 어떻게 최적화할지, (3) 어떤 전사 협업체계가 필요할지 고민해야 함

근거 형태 구체화 필요

- 각 문항에 대한 답변의 근거로 활용할 수 있는 증거물의 형태를 구체화해야 함
- 예를 들어, AI 윤리 실천의 중요 근거로 자주 언급되는 지침·가이드, 평가 방법·기준, 교육 등에 대한 자세한 설명이 도움이 될 것으로 예상됨
- 실제로 시스템 개발 과정 및 모델 작동 방식에 대한 세부 정보(팩트 시트), 검토 조직과 절차 수립 관련 가이드, 이용자 보호 절차 관련 안내 문구 및 약관 등이 어떠한 내용 및 형식을 갖춰야 하는지에 대한 문의가 있었으며, 이에 대한 실질적 대응이 필요하다고 판단됨

3.2 도구의 개발 과정

TF 2차 검토 결과 반영

2023년 10월 31일, 카카오브레인의 파일럿 결과를 TF에 공유하였으며, TF는 11월 9일까지 **도구에 대한 2차 검토(최종 검토)를 진행**하였습니다.

카카오브레인 파일럿 결과에 대한 대응

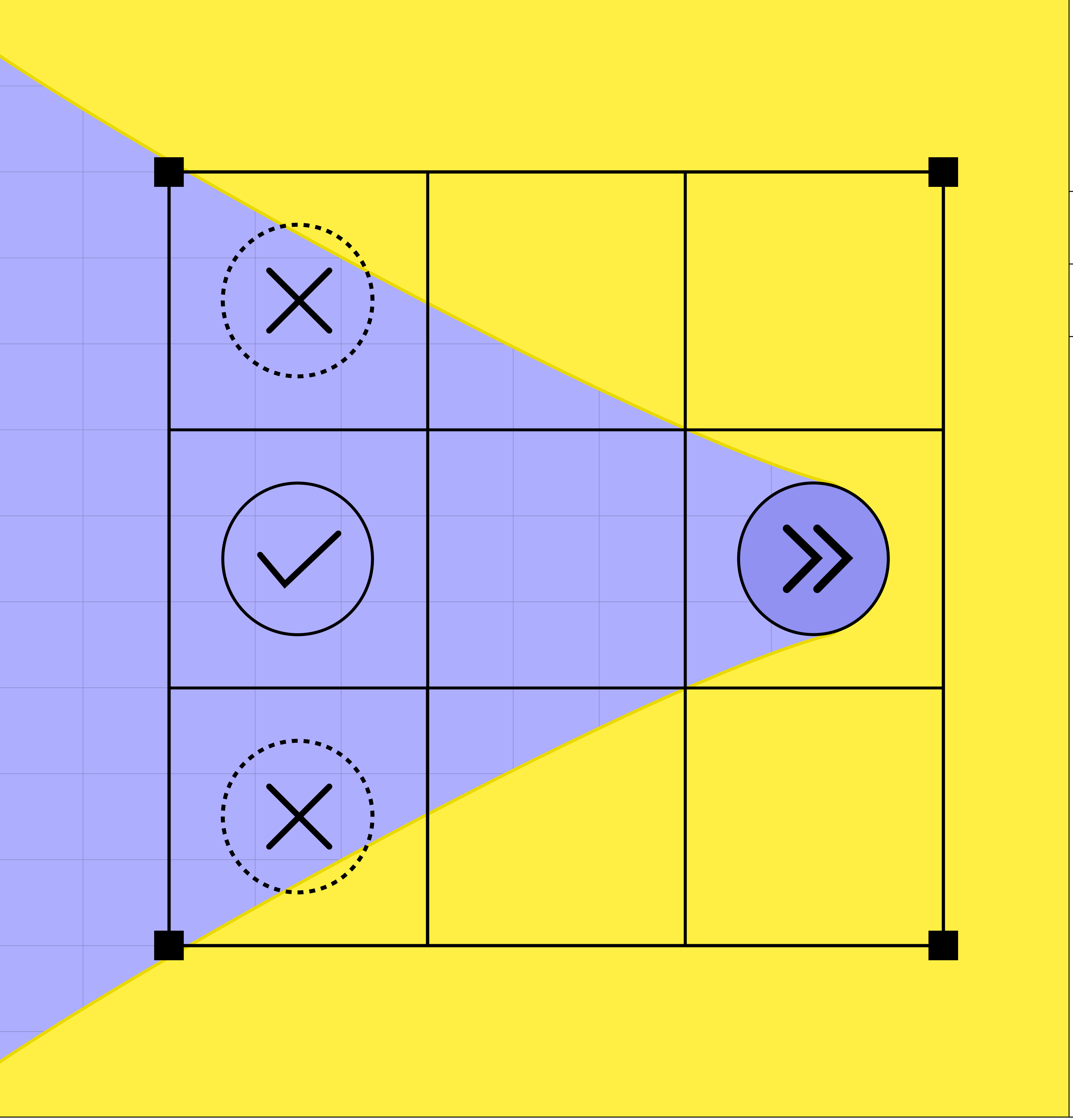
- **문항 수 최소화**: 문항 수가 많을수록 실무자가 자가점검을 시행하기 어려워지는 현실적인 어려움을 고려하여, 총 99문항에서 총 63문항으로 문항 수를 간소화함

TF 2차 검토 기반 방향성 제시

- **도구의 목적 명시**: 사내 AI 윤리 리터러시 향상을 위한 참고 자료인지, 서비스 출시 여부에 영향을 미치는 평가도구인지 등 도구의 목적에 대한 전사 차원의 의사결정이 필요함
- **도구의 실제 적용 가능성 검토**: 해당 도구를 서비스에 실제 적용할 시 어떤 우려가 있는지, 현실적으로 어느 정도로 적용이 가능하고 효과성이 보장되는지에 대한 검토가 필요함
- **문항 간 우선순위 부여**: 책임 있는 AI 확보를 위해서 필수적인 문항에 우선순위를 부여하는 등의 도구의 고도화가 요구되며, 반드시 지켜야 할 원칙이나 사항을 가시화해야 함

내부 회람 및 도구 확정

책임 있는 AI를 위한 생애주기별 자가점검 도구는 카카오와 주요 계열사의 책임 있는 AI 실천을 위한 결의에서 착안되어, 2023년 하반기 카카오 AI 윤리 체크리스트 검토 TF, 카카오 공동체 기술윤리 위원회, 카카오 AI정책지원TF, 카카오 테크직군 위원회, 카카오 서비스비즈직군위원회 등 다양한 사내 이해관계자들의 의견 수렴을 통해 현재의 형태를 갖추게 되었습니다.



개선 방향

4

4.1 도구의 한계

(2023년 말 기준)

현재까지 준비된 **책임 있는 AI를 위한 생애주기별 자가점검 도구**는 향후 실제 서비스 적용의 과정을 여러 번 거침으로써 보다 실용적인 도구로 발전해 가야 합니다.

도구의 효용성 검증

도구의 효용성을 검증하고 개선하기 위해서는 향후 도구 사용자가 될 수 있는 다양한 이해관계자를 대상으로 몇 차례 시범 적용 과정을 거쳐야 합니다. 현재까지 준비된 도구는 별도 TF 및 카카오브레인의 검토를 거쳤지만, 앞으로도 다양한 AI 시스템에 적용됨으로써 도구가 실제 AI 윤리 확보에 어느 정도로 효과적인지 검증해야 합니다.

도구의 객관성 및 신뢰성 확보

도구는 AI 윤리 분야에서 통용되는 참고 자료와 실무진의 경험적 지식을 토대로 개발되었습니다. 하지만 해당 도구가 카카오와 주요 공동체, 더 나아가서 기술 산업의 보편적인 AI 윤리 도구로 발돋움하기 위해서는 외부 기관 또는 전문가의 검토 및 의견 수렴을 통해 객관성 및 신뢰성을 확보할 필요성이 있습니다.

도구의 용도 및 적용 시점 명확화

도구의 용도에 대한 전사 차원에서의 구체적 지침이 필요합니다. 도구를 사내 AI 윤리 리터러시 향상을 위한 참고 자료로 활용할 것인지 또는 서비스 출시 여부에 영향을 미치는 평가도구로 활용할 것인지에 대한 구체적 지침이 필요합니다. 또한, 도구를 실제 적용 시점에 대한 의사결정도 필요합니다.

서비스 맞춤형 도구 개발

도구는 점검 대상인 AI 시스템의 특성과 관계없이 보편적으로 활용될 수 있도록 개발되었습니다. 이러한 이유로 도구를 적용하려는 대상 서비스의 특성에 맞춰 도구를 개선하는 작업이 필요합니다. 이러한 노력을 통해 경험치가 쌓일수록 도구의 활용도는 높아질 것입니다.

도구 고도화

도구의 용도 설정

도구의 구체적인 용도와 기대효과를 설정함으로써 도구의 효용을 극대화하고, 카카오의 AI 윤리 점검 체계 안에 도구를 성공적으로 안착시킬 수 있습니다. 이러한 맥락에서 도구를 (1) 사내 AI 윤리 리터러시 향상을 위한 참고 자료 즉, 관계자가 기획·개발·배포 단계에 있는 AI 시스템의 AI 윤리 준수 정도를 스스로 점검하는 용도로 활용할 것인지 또는 (2) 서비스 출시 여부에 영향을 미치는 평가도구로 활용할 것인지에 대한 전사 차원에서의 의사 결정이 필요합니다.

다양한 AI 서비스 시범 적용

해당 도구는 비단 원칙 또는 이론이 아닌, 즉시 점검에 활용될 수 있는 'AI 윤리 실천 도구'이어야 합니다. 도구를 이러한 목적에 맞게 설계하기 위해서는 최대한 다양한 AI 서비스에 도구를 적용함으로써 실용성을 검증해야 하며, 부족한 부분을 지속해서 도구 설계에 반영해야 합니다.

서비스 맞춤형 도구 개발

각 AI 시스템에 가장 적합한 도구를 개발하기 위해서는 먼저 각 AI 시스템 고유의 특성에 대해 정확히 이해하고 이를 바탕으로 맞춤형 도구를 개발해야 합니다. 예를 들어, 점검하려는 AI 시스템이 AI 모델을 직접 훈련 또는 튜닝시키는지 등에 따라 구체적인 문항을 추가·제거해야 하며, 도구 활용 방식 등에 대한 적절한 설명 및 지침이 제공되어야 합니다. 생성형 AI를 점검할 경우, 학습 데이터의 저작권 문제나 프롬프트 주입 공격 등 생성형 AI에 해당하는 구체적인 고려 사항을 도구에 추가해야 합니다. 더불어, AI 개발 과정에 도구를 사용한 자가점검을 자연스럽게 통합하기 위해서는 문항마다 점검 시행 주체로 적절한 서비스별 점검 담당자를 구체적으로 지정하고, 자가점검 과정과 도구 활용에 관한 정기적인 교육을 제공해야 합니다.

4.2 개선점 제안

(2023년 말 기준)

문항별 우선순위 부여

AI 기술의 급속한 발전에 따라 편리성 등 기회가 많아졌지만 동시에 다양한 사회적 문제 또한 대두되고 있습니다. 카카오는 근본적인 기술 윤리적 고민에 더불어 당장 이용자와 사회에 큰 영향을 미칠 수 있는 기술 윤리 문제들에 대해서도 신속하게 대응해야 합니다. 즉, AI 산업과 기술의 사회적 '영향 범위,' '규모,' '회복 가능성' 등을 고려하여 도구의 각 문항에 우선순위를 부여해야 합니다.

정성 및 정량 결과 도출

도구를 통한 자가점검 결과가 정성 및 정량적 결과를 균형 있게 보여줄수록 향후 AI 윤리 실천 방향을 설정하는 데 유리합니다. 현재까지 보완된 도구도 문항별로 상응하는 AI 윤리 원칙을 표시해 두어, 문항을 답하며 AI 윤리 원칙 준수 여부 및 정도를 알 수 있습니다. 하지만, 도구 활용 결과의 정량화 방식 및 정도에 대한 고도화 작업이 더 진행되어야 하며, 어떤 기대효과를 염두에 두고 정성 문항을 추가할지도 고민해 보아야 합니다. 예를 들어, AI 윤리 점검 대상인 AI 시스템에 대한 균형 잡힌 평가를 위해서는 해당 시스템의 부정적 영향뿐만 아니라 긍정적 효과에 대한 정성 평가 문항도 추가해야 합니다. 정량 및 정성 데이터를 아우르는 종합 결과지를 도구 사용자에게 제공한다면 자가점검 결과를 토대로 향후 계획을 수립하는 데 보다 실질적 도움이 될 것입니다.

외부 기관·전문가 자문

외부 기관·전문가 자문은 도구의 객관성 및 신뢰성을 확보하는 동시에 외부 전문가의 관점을 반영하여 AI 윤리 실천에 새로운 방향성을 제시하기도 합니다. 학계 전문가와의 협력은 실천 중심인 카카오의 AI 윤리 활동이 전문적 지식을 토대로 하며 그 논리와 체계를 개선하는 계기가 될 수 있습니다. 또한 국제기구와의 협력을 통해 AI 윤리와 관련된 국제 규범 또는 가이드라인을 국내에 소개하는 등 다각도에서 카카오의 영향력을 넓히는 데 기여할 수 있습니다.

운영 체계화

AI 윤리 원칙의 지속적 발전에 대한 공약

AI 윤리 실천에 대한 기업의 책임을 경영진이 앞장서서 공식적으로 선언하는 것은 지대한 영향력을 가집니다. 조직 내 경영자, 임직원 등 고위직이 카카오가 제공하는 AI 서비스의 품질, 특히 사회적 영향력을 진정성 있게 고민하고, 숙고의 과정에 다양한 이해관계자를 합류시키며, AI 윤리에 대한 전사 차원의 비전 및 중장기적 계획을 대내외로 공표하는 것은 책임 있는 AI 실천을 위한 카카오의 의지 표명이자 긴 과정의 첫걸음이 될 것입니다.

AI에 대한 통합적 통제 및 관리 방안

AI 윤리가 고려되는 사내 절차를 구축하기 위해서는 우선 전사 차원에서 합의한 'AI 통합 통제 및 관리 방안'이 마련되어야 합니다. 현재 카카오가 기획·개발하는 AI 서비스는 카카오 AI정책지원TF에 전달하게 되어있습니다. 해당 TF는 전달받은 AI 서비스를 다양한 관점(개인정보, 법무, 대외, 정책, 인권과 기술 윤리 등)에서 검토하여 리스크 검토하고 대응 방안을 제시하는 등의 역할을 하고 있습니다. AI 서비스별 대응도 중요하지만, 앞으로 보다 장기적이고 통합적인 AI 정책 적용이 필요합니다.

'AI 시스템을 계획하고 설계'하는 시점부터 '데이터를 수집하고 처리'하는 과정을 거쳐, 'AI 모델을 개발'하고, '시스템을 기획, 구현'한 후, 마지막으로 '운영 및 모니터링'하는 단계까지 전사적으로 AI 시스템과 관련된 다수의 이해관계자가 유기적으로 업무 협조를 할 수 있는 체계를 구축해야 합니다. 특히, 도구의 실제 활용을 대비하여, '카카오 AI정책지원TF,' '카카오 공동체 기술윤리 위원회,' 'ESG 총괄 회의' 등 다양한 계위의 협의체 간 역할 및 협조 요청 순서 등이 구체화하여야 합니다.

구체적 실행 계획 및 성과 모니터링

도구에 대한 외부 자문, 여러 서비스를 대상으로 한 도구의 시범 적용, 도구에 대한 사내 교육 등 카카오는 더욱 구체적인 AI 윤리 실행 계획을 바탕으로 관련 성과를 지속해서 모니터링하고, 이를 바탕으로 도구 자체의 발전, 더 나아가 카카오의 AI 윤리 거버넌스 고도화에 기여할 수 있도록 해야 합니다.

5.1 참고 문헌

- **2022 신뢰할 수 있는 인공지능 개발 안내서(안)**
  (2022. 과학기술정보통신부, 한국정보통신기술협회)
- **2023 인공지능 윤리기준 실천을 위한 자율점검표**
  (2023. 과학기술정보통신부, 정보통신정책연구원)
- **인공지능 인권영향평가 도입 방안연구**
  (2022. 국가인권위원회)

5.2 녹서 집필진

기획·편집

카카오 CA협의체 ESG위원회 그룹기술윤리팀

검토

카카오 CA협의체 ESG위원회 그룹 기술윤리 소위원회

카카오 AI정책지원 TF

카카오 테크직군위원회

카카오 서비스비즈직군위원회

카카오 AI윤리 체크리스트 검토 TF

5.3 녹서 발행일 및 발행처

발행연월

2024년 7월 31일

발행처

카카오 CA협의체 ESG위원회 그룹 기술윤리 소위원회

기획·편집 문의처

카카오 CA협의체 ESG위원회 그룹기술윤리팀

디자인

매뉴얼

kakao