인공지능을 위한 일상의 윤리

디자이너와 개발자를 위한
실용 가이드

IBM

2018년 9월

Adam Cutler: IBM 특별 디자이너
Milena Pribić: 디자이너
Lawrence Humphrey: 디자이너

감사의 말

Francesca Rossi - AI 윤리 부문 글로벌 리더, IBM Research 우수 직원
Anna Sekaran - IBM 프로그램 리더, 코그너티브 컴퓨팅, 학술 사업 및 홍보 담당
Jim Spohrer – IBM Cognitive OpenTech 담당 이사
Mike Monteiro – Mule Design의 공동 설립자 겸 디자인 담당 이사
Zack Causey: 디자이너/일러스트레이터

목 차

문서의 목적

"제도 테이블에서 지우개를 사용하거나 건설 현장에서 망치를 사용할 수 있습니다."

– Frank Lloyd Wright

이 문서는 AI에 대한 일상의 윤리를 정의하고자 하는 대화의 시작을 나타냅니다. 윤리는 AI의 시작 단계부터 디자인 및 개발 프로세스에 포함되어야 합니다.

이 문서는 아이디어를 자극하고 사고를 촉진하기 위한 것입니다. 여기에서 아이디어는 **단순하게 시작하여 반복**하는 것입니다.

처음부터 완벽을 추구하기보다는 이 문서를 읽고 사용하는 모두가 의견을 내고 비평하고 반복되는 이후의 작업에 참여할 수 있도록 하기 위해 이 간행물을 발간합니다. **그러니 여기에서 알게 된 것을 실험하고 시도하고 사용해 보고 피드백을 보내주십시오.**

AI 시스템 디자이너와 개발자는 이러한 개념을 인지하고 **이러한 아이디어를 의도적으로 실천**할 수 있는 기회를 잡아 활용하시길 권장해 드립니다.

팀 및 다른 사람들과 함께 작업할 때 이 가이드를 공유하시기 바랍니다.

질문, 의견 또는 제안이 있는 경우 edethics@us.ibm.com으로 이메일을 보내 이러한 노력에 동참하시기 바랍니다.

Adam Cutler

IBM 특별 디자이너,
인공지능 디자인

Milena Pribić

IBM 디자이너,
인공지능 디자인

Lawrence Humphrey

IBM 디자이너,
인공지능 디자인

최신 정보 받기! 최신 버전을 다운로드하려면 여기를 클릭하십시오.

개요

윤리적 의사결정은 기술 문제 해결의 또 다른 형태가 아닙니다

AI 디자이너와 개발자로서 우리는 집단적 영향력의 큰 부분을 차지하고 있습니다. 우리는 수백만 명의 사람에게 영향을 미칠 시스템을 만들고 있습니다.

인공지능 기술은 기능과 영향 면에서 빠르게 확장되고 있습니다. AI 시스템의 디자이너와 개발자로서 우리는 작업의 윤리적 고려 사항을 이해해야 합니다.

지능형 시스템의 기능 향상에만 집중된 기술 중심의 관점은 사람의 요구를 충분히 고려하지 못합니다.

AI가 영향을 미치는 사회 또는 커뮤니티의 가치 및 윤리적 원칙과 일치하는 방식으로 윤리적인 인간 중심의 AI가 디자인되고 개발되어야 합니다.

윤리는 권리, 의무, 사회적 혜택, 공정성 또는 특정 덕목(Markkula Center for Applied Ethics[1])의 관점에서 인간이 무엇을 해야 하는지를 규정하는 잘 정립된 옳고 그름의 기준을 기반으로 합니다.

이 가이드는 다음과 같은 토론 주제를 제시합니다.

a) AI 시스템이 갖추어야 할 특정 덕목

b) AI를 구축하고 교육하는 디자이너와 개발자를 위한 지침

인간과 기계 간에 신뢰를 생성하고 육성하기 위해서는 AI의 디자인, 구축 및 유지 관리 시 참조할 수 있는 윤리 관련 리소스와 표준을 이해해야 합니다.

이 문서 전반에서 언급될 자율 및 지능형 시스템의 윤리에 대한 IEEE 글로벌 이니셔티브[2]와 같은 그룹의 AI 윤리에 대한 지대한 관심은 모든 규모의 비즈니스와 작업 그룹에 반영되어야 합니다.

윤리적 AI 시스템의 기준과 지표는 결국 해당 시스템이 운영되는 산업과 사용 사례에 따라 달라질 것입니다.

우리는 이 문서가 팀이 모범 사례를 구축하는 데 도움이 되는 중요한 자료가 되길 바랍니다.

디자이너와 개발자는 고립된 상태로 일해서는 안 되고 사용자의 요구와 우려에 귀를 기울여야 합니다.

디자인 및 개발팀에서 사용자의 우려를 해결하도록 하려면 지속적인 개선과 평가가 중요합니다. 이 문서는 팀에 출발점을 제공하며 AI 기능이 계속 발전함에 따라 이 문서도 함께 발전할 것입니다.

5가지 주요 윤리적 영역

01. 책임
02. 가치 정렬
03. 의사결정 과정 이해 능력
04. 공정성
05. 사용자 데이터 권한

시간이 지나며 AI 기능이 확장됨에 따라 이러한 주요 윤리적 영역을 이해하고 발전시키는 것은 우리 공동의 책임입니다. 이러한 주요 영역은 AI 시스템의 구축 및 사용을 위한 윤리적 기반을 확립하기 위한 의도적인 프레임워크를 제공합니다.

AI 디자이너와 개발자는 "디자인에서의 인공지능 및 윤리[3]"이라는 IEEE 과정에서 다루는 주제를 좀 더 깊이 탐구하고 싶을 수 있습니다. 이 과정에서는 (1) AI 시대의 책임 있는 혁신: 철학적 기반, 이익 및 사회적 목적으로 AI를 사용하는 회사, (2) 비즈니스를 위한 윤리적 디자인의 경제적 이점: 지능형 시스템, 윤리 및 정부 정책, (3) 알고리즘 시대에서 디자인의 가치: 도덕적, 사회적 및 법적 가치를 식별, 분석 및 실천, (4) 넛징의 본질: 사람에게 영향을 미치는 AI 기능은 선 또는 악을 위해 사용될 수 있다, (5) 데이터 보호 및 데이터 안전: 일반 데이터 보호 규정 및 AI 시스템 구축 및 유지 관리에서 데이터의 중요성 등을 다룹니다.

또한, IBM Research 팀은 서비스의 투명성을 높이고 신뢰를 쌓을 수 있도록 AI 서비스 개발자 및 공급자가 **공급자 적합성 선언**(Supplier's Declaration of Conformity, SDoC 또는 팩트시트)[4]을 자발적으로 완성하고 발표할 것을 제안했습니다. 식품용 영양 레이블이나 가전제품 정보 시트와 같이 AI 서비스의 팩트시트는 제품의 중요한 특징에 대한 정보를 제공할 것입니다. 주요 영역과 함께 이 제안을 더욱 발전시켜 신뢰할 수 있는 AI 시스템의 시대를 열고 폭넓은 도입을 촉진할 수 있기를 바랍니다.

실행 예제

한 호텔 체인은 객실 가상 비서/컨시어지에 인공지능을 탑재하여 사용자의 투숙을 지원하고 개인화된 서비스를 지원합니다. 담당 프로젝트팀을 이 문서 전반에서 예제로 사용하겠습니다. 이 대화 에이전트에는 다음과 같은 기능이 포함됩니다.

- 실제 상담원과 같은 가상 서비스
- 투숙객이 선호하는 언어로 지원되는 객실 서비스
- 자연어로 지원되는 객실 제어
- 객실 가상 비서를 통해 서비스팀에 직접 요청 전달

01.

책임

AI 디자이너와 개발자는 AI 디자인, 개발, 의사결정 과정 및 결과를 고려해야 할 책임이 있습니다

인간의 판단은 객관적으로 보이는 논리적 의사결정 시스템 전반에서 역할을 수행합니다. 알고리즘을 작성하는 것도, 성공과 실패를 정의하는 것도, 시스템 사용에 대한 결정을 하는 것도, 시스템의 결과에 영향을 받을 수 있는 것도 모두 인간입니다.

어떤 단계에서든 AI 창조에 관여한 모든 사람은 개발에 투자한 기업과 마찬가지로 그 시스템이 세상에 미치는 영향을 고려할 책임이 있습니다.

권장 조치:

1. 업무 또는 책임에 관한 문제에 대해 혼란스러워하는 사람이 없도록 처음부터 설계팀과 개발팀이 회사 정책을 명확히 이해하고 액세스할 수 있게 합니다. AI 디자이너 또는 개발자로서, **아는 것은 여러분의 책임입니다.**

2. 회사/소프트웨어의 책임이 어디에서 끝나는지 이해합니다. 사용자, 클라이언트 또는 기타 외부 소스가 데이터나 도구를 사용하는 방식을 제어할 수 없을 수 있습니다.

3. 디자인 프로세스와 의사결정에 대한 상세한 기록을 남깁니다. 모범 사례와 반복을 장려할 수 있도록 디자인 및 개발 프로세스에서 기록을 유지하기 위한 전략을 결정합니다.

4. 귀사의 비즈니스 행동 지침을 준수합니다. 또한, AI가 준수해야 하는 국내 및 국제법, 규정 및 지침[5]을 이해합니다. 기타 관련 리소스는 IEEE 윤리적 디자인 문서[6]에서 확인할 수 있습니다.

"설문 조사에 참여한 개발자 중 거의 50%가 AI를 만드는 사람이 기술의 파급 효과를 고려해야 한다고 답했습니다. 상사나 중간 관리자가 아니라 프로그래머가 말입니다."

- Mark Wilson, Fast Company[7], 2018년 Stack Overflow의 개발자 설문 조사 결과[8]

고려 사항:

1. 개인적으로 알고리즘을 개발하고 모니터링하지 않더라도 AI의 작동 방식을 이해합니다.

2. 전체론적 맥락에서 윤리 문제를 이해할 수 있도록 사회학자, 언어학자, 행동학자 및 기타 전문가의 이차적 연구를 참조합니다.

팀에 질문할 내용:

1. AI 시스템에 대한 사용자 영향 수준에 따라 책임은 어떻게 변할까요?

2. AI를 인간의 의사결정 과정에 포함시켜야 할까요, AI 스스로 의사결정을 내리게 할까요, 아니면 이 둘을 혼합해야 할까요?

3. 팀에서는 프로세스에 대한 기록을 어떻게 남겨야 할까요?

4. AI 시스템 가동 후 윤리적 디자인 선택과 고려 사항을 어떻게 추적할까요?

5. 우리 작업을 처음 접하는 사람들이 우리 기록을 이해할 수 있을까요?

책임의 예

- 팀에서 디자인 연구원을 활용하여 호텔의 실제 투숙객과 대면 인터뷰를 통해 고객의 바람과 요구를 파악합니다.

- 호텔 가상 비서의 응대가 투숙객의 요구나 기대에 미치지 못하는 경우 이를 팀의 책임으로 여깁니다. 선호 사항을 더 파악하기 위해 고객의 피드백을 반영하는 체계를 구축했고 투숙객이 원할 경우 언제든 AI 서비스를 종료하도록 했습니다.

02.

가치 정렬

AI를 설계할 때 사용자 그룹의 규범과 가치를 염두에 두고 이에 부합하도록 해야 합니다.

AI는 다양한 인간 관심사와 함께 작동합니다. 사람들은 경험, 기억, 가정교육 및 문화적 규범을 비롯하여 다양한 맥락 요인을 기반으로 의사결정을 내립니다. 이러한 요인들로 인해 우리는 가정, 사무실 또는 어디서든 포괄적인 맥락에서 "옳고 그름"을 근본적으로 이해할 수 있습니다. 우리는 의지할 수 있는 풍부한 경험이 있으므로 이를 인간의 제2의 천성으로 보기도 합니다.

오늘날 AI 시스템에는 이러한 유형의 경험이 없으므로 디자이너와 개발자는 기존 가치를 고려할 수 있도록 서로 협력해야 합니다. 다양한 문화적 규범과 가치를 세심하게 살필 수 있도록 주의해야 합니다. 가치 체계를 고려하는 것이 어려워 보일 수 있지만, 보편적 원칙의 공통적인 핵심은 그것이 협동 현상이라는 것입니다. 성공적인 팀은 협력과 협업이 최상의 결과를 가져온다는 것을 이미 이해하고 있습니다.

권장 조치:

1. 설계하고 있는 가치 체계를 확립하는 문화에 대해 고려합니다. 가능하면 팀에서 관련 관점을 명확히 표현하는 데 도움을 줄 수 있는 정책 담당자 및 학자를 초빙합니다.

2. 디자인 연구원과 협력하여 사용자의 가치를 이해하고 반영합니다. 이 프로세스에 대한 자세한 내용은 여기를 참조하십시오[9].

3. Ethics Canvas[10]를 사용하여 사용자의 가치에 대한 이해를 지도로 표시하고 그에 따라 AI의 동작을 조정하는 것을 고려합니다. 가치는 특정 활용 사례와 해당 가치의 영향이 미치는 커뮤니티에 특유한 것입니다. 조정을 통해 사용자는 AI의 동작과 의도에 대한 이해를 높일 수 있습니다.

"머신이 자율적 주체로서 인간 공동체에 참여한다면, 이러한 주체에게도 공동체의 사회적 및 도덕적 규범을 따를 것을 기대하게 됩니다.

머신이 그렇게 할 수 있도록 하는 데 필요한 단계는 해당 규범을 파악하는 것입니다. 하지만 누구의 규범을 말하는 걸까요?"

- 자율적 및 지능형 시스템의 윤리에 대한 IEEE 글로벌 이니셔티브[11]

고려 사항:

1. 어디에서 시작해야 할지 모르겠다면 IBM의 기업 책임 표준[12]을 검토하거나 귀사의 표준 문서를 활용하십시오.

2. 가치는 주관적이고 전세계적으로 차이가 있습니다. 따라서, 글로벌 기업은 언어 장벽과 문화 차이를 고려해야 합니다.

3. 선의의 가치가 의도하지 않은 결과를 초래할 수 있습니다. 예를 들어 맞춤형 정치 뉴스 피드는 신념에 맞는 뉴스를 제공하지만 전체론적 게슈탈트를 나타내지는 않습니다.

팀에 질문할 내용:

1. 우리 AI에서는 어떤 그룹의 가치를 표현하며 그 이유는 무엇인가요?

2. 팀으로서 고려해야 할 가치에 대한 합의를 도출하려면 어떻게 해야 하나요?(도덕적 합치에 대한 자세한 내용은 여기를 참조하십시오[13])

3. 시간이 지나면서 가치가 진화함에 따라 우리 AI에 반영된 가치를 변경하거나 조정하려면 어떻게 해야 하나요?

가치 정렬의 예

- 팀은 음성 인식 비서가 제대로 작동하려면 AI를 작동시키는 명령어를 "항상 듣고" 있어야 한다는 것을 이해하고 있습니다. 팀에서는 AI 호텔 가상 비서가 AI를 작동시키는 명령어를 듣고 있지만 투숙객이 알지 못하는 상태에서 데이터를 유지하거나 투숙객을 모니터링하도록 설계하지 않았다는 것을 투숙객에게 명확히 설명합니다.

- AI를 작동시키는 명령어를 듣는 중에 수집된 오디오는 5초마다 자동 삭제됩니다. 투숙객이 AI 서비스 활용에 동의했다 하더라도, AI를 작동시키는 명령어를 부르지 않는 한 투숙객의 소리를 적극적으로 듣지 않습니다.

- 팀은 이 AI 에이전트가 전 세계 호텔에서 사용될 것이므로 서로 다른 언어와 관습이 필요하다는 것을 알고 있습니다. 팀은 언어학자와 논의하여 AI가 투숙객의 언어로 말하고 해당 관습을 존중하도록 합니다.

03.

설명 가능성

AI를 설계할 때 사람이 AI의 의사결정 과정을 쉽게 인지, 감지 및 이해할 수 있도록 해야 합니다.

일반적으로 우리는 논리적인 설명을 못하는 사람을 맹목적으로 신뢰하지는 않습니다. 이는 AI도 마찬가지이며 오히려 더 심할 것입니다.[14] AI의 능력이 향상되고 영향 범위가 확장됨에 따라 AI의 의사결정 과정을 사람이 이해할 수 있는 방식으로 설명할 수 있어야 합니다. 이처럼 설명 가능성은 AI와 상호 작용하는 사용자가 AI의 결론과 추천을 이해하는 데 매우 중요합니다. 좋은 설계는 원활한 경험을 위해 투명성을 희생시키지 않습니다. **지각할 수 없는 AI는 윤리적 AI가 아닙니다.**

권장 조치:

1. 질문을 허용합니다. 사용자가 AI가 무엇을 왜 하고 있는지 지속적으로 물어볼 수 있어야 합니다. 이는 사용자 인터페이스에서 항상 명확하고 솔직해야 합니다.

2. 특히 AI가 개인 식별 정보, 개인 건강 정보 및/또는 생체 정보와 같은 매우 민감한 개인 정보 데이터를 다루는 경우, 의사결정 과정을 검토할 수 있어야 합니다.

3. AI의 도움을 받아 사용자가 매우 민감한 의사결정을 내리는 경우, AI는 사용자에게 추천에 대한 충분한 설명, 사용된 데이터 및 추천의 근거를 제공할 수 있어야 합니다.

4. 팀은 AI의 의사결정 과정에 대한 기록에 액세스하고 이러한 의사결정 과정을 검증할 수 있어야 합니다.

"IBM은 AI 시스템이 어떻게 결론이나 추천에 도달하게 되었는지 사람들이 이해할 수 있도록 투명성 및 데이터 거버넌스 정책을 지원합니다. 기업은 자사의 추천 알고리즘이 어떻게 작동하는지 설명할 수 있어야 합니다. 만약 할 수 없다면 그러한 시스템은 시장에 출시되어서는 안 됩니다."

- IBM의 데이터 책임[15]

고려 사항:

1. 혁신적인 기술(disruptive technology)에서 대중의 신뢰를 구축하고 안전한 관행을 촉진하며 광범위한 사회적 도입을 지원하기 위해서는 의사결정 과정을 설명하는 능력(explainability)이 필요합니다.

2. *예를 들어 금융 투자 알고리즘*과 같이 AI가 수행하는 전체 의사결정 과정에 사용자가 액세스할 수 없는 상황이 있습니다.

3. AI 시스템의 투명성 수준이 아무런 흠 없이 유지되고 있는지 확인합니다. 사용자는 세부적인 AI 프로세스에 액세스할 수 없더라도 AI의 의도에 대한 일반적인 정보를 알 수 있어야 합니다.

팀에 질문할 내용:

1. 사용자 경험을 저해하거나 당면한 작업에 방해가 되지 않으면서 의사결정 설명 능력을 경험에 통합하려면 어떻게 해야 할까요?

2. 보안 또는 IP 상의 이유로 특정 프로세스나 정보를 사용자로부터 숨길 필요가 있나요? 사용자에게 이 점을 어떻게 설명하나요?

3. AI 의사결정 프로세스 중 쉽게 이해하고 설명할 수 있는 방법으로 사용자에게 표현할 수 있는 부분은 무엇일까요?

설명 가능성의 예

- GDPR(일반 데이터 보호 규정)[16]에 따라 투숙객은 호텔 객실의 가상 비서를 사용할지 여부를 명시적으로 선택해야 합니다. 또한, AI가 어떻게 추천하고 제안하는지를 보여주는 투명한 UI를 제공해야 합니다.

- 팀의 연구원은 인터뷰를 통해 호텔 투숙객이 개인 정보 저장 여부를 선택할 수 있는 기능을 원한다는 것을 알고 있습니다. 팀에서는 AI가 투숙객에게 음성 또는 그래픽 UI를 통해 옵션을 제공하고 시스템이 동의 하에 정보를 수집할 수 있는 기능을 제공합니다.

- 허가를 받은 상태에서 AI는 체류하는 동안 방문할 곳을 추천합니다. 투숙객은 이러한 추천을 하게 된 이유와 추천을 위해 어떤 데이터 세트가 활용되고 있는지 물어볼 수 있습니다.

04.

공정성

AI는 편향을 최소화하고 포용적 표현을 장려하도록 설계해야 합니다.

AI는 개인의 민감한 데이터를 다루게 될 경우 사생활에 대한 더욱 깊은 인사이트를 제공합니다. 인간은 본질적으로 편향에 취약하고 AI를 구축하는 것도 인간이므로 우리가 만드는 시스템에 편향이 포함되어 있을 가능성이 있습니다. 다양한 인구를 대표하는 지속적인 연구와 데이터 수집을 통해 알고리즘의 편향을 최소화하는 것이 담당 팀의 역할입니다.

권장 조치:

1. AI 실시간 분석은 의도적인 편향과 의도하지 않은 편향을 동시에 드러냅니다. 데이터 편향이 명백해지면 팀에서는 해당 데이터의 출처와 완화 방법을 조사하고 파악해야 합니다.

2. 디자인 및 개발을 할 때 의도적인 편향이 없도록 하고, 팀 리뷰를 통해 의도하지 않은 편향을 방지합니다. 의도하지 않은 편향에는 고정 관념, 확증 편향 및 매몰 비용 편향이 포함될 수 있습니다(페이지 26 참조).

3. 피드백 메커니즘을 가동하거나 사용자와의 대화를 통해 사용자가 확인하는 편향 또는 문제에 대한 인식을 높입니다. *예를 들어 Woebot[17]은 링크를 제안한 후 "어떻게 생각하는지 알려주십시오"라고 묻습니다.*

"AI를 위한 새로운 윤리적 프레임워크를 발전시키고 데이터 세트의 품질과 인간이 어떻게 AI를 인식하고 작업하는지에 대해 비판적으로 생각함으로써 우리는 모두에게 이익에 되는 방식으로 [AI] 분야를 가속화할 수 있습니다. IBM은 [AI]가 AI 시스템의 편향을 완화할 수 있는 열쇠를 쥐고 있으며 우리가 인간으로서 가지고 있는 기존의 편향을 해결할 수 있는 전례 없는 기회를 제공한다고 믿습니다."

- AI의 편향: 공정한 AI 시스템을 구축하고 편향성을 완화하는 방법[18]

고려 사항:

1. 팀이 다양하면 편향을 최소화할 수 있는 좀 더 다양한 경험을 표현하는 데 도움이 됩니다. 다양한 나이, 민족, 성별, 교육 분야, 문화적 관점을 가진 팀원을 받아들입니다.

2. AI는 수집하는 데이터 유형에 따라 다양한 유형의 편향에 취약할 수 있습니다. 문제에 신속하게 대응할 수 있도록 교육과 결과를 모니터링하고, 일찍 그리고 자주 테스트합니다.

팀에게 질문할 내용:

1. AI 시스템의 디자인 및 개발 중에 발생하는 의도하지 않은 편견을 식별하고 감사하려면 어떻게 해야 할까요?

2. 현재 상황은 시간이 지나면서 변합니다. 진행 중인 데이터 수집 작업에 이러한 변화를 반영할 수 있는 방법을 적용하려면 어떻게 해야 할까요?

3. 디자인 또는 의사결정에서 의도하지 않은 편향을 바로잡기 위해 사용자의 피드백을 수집하는 가장 좋은 방법은 무엇일까요?

공정성의 예

- 호텔의 글로벌 관리팀과 자리를 같이 한 후, 팀에서는 다양성과 포괄성이 호텔의 가치에 중요한 요소라는 사실을 알게 되었습니다. 그 결과, 팀에서는 사용자의 인종, 성별 등에 대해 수집된 데이터를 AI 사용과 결합하여 특정 인구 집단을 대상으로 마케팅하거나 그들을 배제하는 데 사용하지 않도록 했습니다.

- 팀에서 호텔의 투숙객에 대한 데이터 세트를 전달받았습니다. 이 데이터를 분석하여 에이전트 구축에 활용한 후, 데이터에 어느 정도 알고리즘적 편향이 있다는 것을 알게 됩니다. 팀에서는 더 크고 더 다양한 데이터 세트로 모델을 교육하는 데 시간을 할애합니다.

의도하지 않은 편향의 정의

일반적인 지식 근로자는 다양한 유형의 편향에 대해 인지하지 못합니다. 다음의 목록이 전부는 아니지만 이러한 편향은 AI를 디자인하고 개발할 때 좀 더 주의해야 하는 유형에 포함됩니다.

지름길 편향 (shortcut bias)

"나는 이것에 대해 생각할 시간이나 에너지가 없다."

가용성 편향(availability bias)

기억 속의 더 큰 "가용성"으로 이벤트를 과대 평가 – 최근, 특이한 또는 감정이 북받친 기억의 영향을 받을 수 있음.

기저율 무시 오류(base rate fallacy)

일반 정보를 무시하고 특정 정보(특정 사례)에 집중하는 경향.

부합성 편향(congruence bias)

대체 가설을 테스트하는 대신 직접 테스트를 통해 가설을 배타적으로 테스트하는 경향.

공감 격차 편향(empathy gap bias)

자신 또는 다른 사람의 감정의 영향이나 힘을 과소평가하는 경향.

고정관념(stereotyping)

그룹에 속한 구성원에 대한 실제 정보 없이 해당 개인에게 특정 특성이 있을 것으로 예상.

공평성 편향 (impartiality bias)

"내가 가끔 틀리는 건 알지만, 이건 내가 옳다"

앵커링 편향(anchoring bias)

의사결정을 내릴 때 하나의 특성이나 일부 정보에 너무 많이 의존(일반적으로 해당 주제에 대해 확보한 첫 번째 정보).

시류 편향(bandwagon bias)

많은 사람이 하기 때문에 무언가를 하거나 믿는 경향(그룹 사고).

편향 사각지대(bias blind spot)

자신을 다른 사람보다 덜 편향된 것으로 보거나 자신보다 다른 사람에게서 더 많은 인지적 편향을 발견하는 경향.

확증 편향(confirmation bias)

자신의 선입관을 확인하는 방법으로 정보를 검색하거나 해석하고 그러한 정보에 집중하는 경향.

후광 효과(halo effect)

전반적인 인상이 관찰자에게 영향을 미치는 경향. 한 영역의 긍정적인 감정이 모호하거나 중립적인 특성을 긍정적으로 보게 함.

자기중심적 편향 (self-interest bias)

"우리가 가장 많은 기여를 했습니다. 그들은 별로 협조적이지 않았습니다."

내집단/외집단 편향(ingroug/outgroup bias)

자신이 속하지 않은 집단보다 자신이 속한 집단에 더 호의적인 경향.

매몰 비용 편향(sunk cost bias)

비록 더는 유효하지 않아 보여도 과거의 선택을 정당화하려는 경향.

현상 유지 편향(status quo bias)

더 나은 대안이 존재하는 경우에도 현재 상황을 유지하려는 경향.

NIH(여기에서 만든 것이 아님) 편향(Not Invented Here bias)

외부 그룹에서 개발된 제품, 연구, 표준 또는 지식과 접촉하거나 사용하는 것에 매우 배타적.

자기 위주 편향(self-serving bias)

강점/업적에 집중하고 단점/실패를 간과하는 경향. 자신이 속한 그룹의 작업에 대한 책임을 다른 그룹에 더 많이 떠 넘기는 것.

05.

사용자 데이터 권한

AI는 사용자 데이터를 보호하고 액세스 및 사용에 대한 사용자의 권한을 보존하도록 설계되어야 합니다.

사용자에게 상호 작용을 제어할 수 있는 권한을 부여하는 것은 팀의 책임입니다.

Pew Research[19]의 최근 조사에 따르면 미국인 74%가 자신의 정보에 대한 제어권을 유지하는 것이 "매우 중요하다"고 밝혔습니다. 유럽연합 집행위원회[20]는 EU 시민의 71%가 기업이 자신의 허락 없이 자신에 대한 정보를 공유하는 것은 용납할 수 없다고 생각하는 것을 확인했습니다. 이러한 비율은 AI가 개인정보 보호를 강화하거나 침해하는 데 사용됨에 따라 증가할 것입니다. 사용자들이 AI가 자신의 이익을 위해 사용된다는 것을 이해하게 하려면 EU 일반 데이터 보호 규정[21]에서 적용되는 부분과 다른 국가에서 유사한 규정을 모두 준수해야 합니다.

권장 조치:

1. 사용자는 어떤 데이터가 어떤 맥락에서 사용되는지 언제나 제어할 수 있어야 합니다. 사용자는 AI가 알거나 사용하면 손상되거나 부적합할 수 있는 개인 데이터에 대한 액세스를 거부할 수 있습니다.
2. AI가 상호 작용하기 전에 허가를 요청하도록 하거나 상호 작용 중에 옵션을 제공하여 사용자가 서비스 또는 데이터를 거부할 수 있도록 합니다. 개인정보 설정 및 권한은 명확하고 검색 가능하며 조정 가능해야 합니다.
3. 개인 정보를 사용하거나 공유하는 방식을 완전히 공개합니다.
4. 사용자의 데이터는 도난, 오용 또는 데이터 손상으로부터 보호되어야 합니다.
5. 새로운 AI 서비스를 생성할 때 허가 없이 다른 회사의 데이터를 사용하는 것을 금지합니다.
6. AI의 허용 가능한 사용자 데이터 액세스 권한을 설계할 때 해당하는 국내 및 국제 권리 법률[22]을 준수합니다.

"개인이 개인 정보의 묶음판매(bundling) 또는 재판매의 결과를 명시적으로 알 수 있도록 정책 및 관행과 함께 고유한 자격 증명과 개인 데이터를 관리할 수 있는 메커니즘이 필요합니다."

- 자율 및 지능형 시스템의 윤리에 대한 IEEE 글로벌 이니셔티브[23]

고려 사항:

1. 암호화, 액세스 제어 방법론, 동의 관리 모듈을 비롯한 보안 관행을 적용하여 권한이 있는 사용자로 액세스를 제한하고 사용자 기본 설정에 따라 데이터에서 개인 정보를 제거합니다.

2. 팀과 협력하여 이러한 보안 관행에서 부족한 부분을 해결하는 것은 디자이너와 개발자 여러분의 책임입니다.

팀에 질문할 내용:

1. AI가 활용하는 민감한 개인 데이터의 유형은 무엇이며 이러한 데이터는 어떻게 보호됩니까?

2. 데이터 사용에는 어떤 계약이 필요하며 AI에 적용되는 현지 및 국제법은 무엇입니까?

3. 최소한의 필수 사용자 데이터로 최상의 사용자 경험을 만들려면 어떻게 해야 하나요?

데이터 권한의 예

- 호텔은 투숙객이 AI 서비스를 이용하기 전에 AI 호텔 비서 사용에 대한 동의서를 제공합니다. 이 동의서에서는 호텔 측이 데이터를 소유하지 않으며 투숙객은 언제든지, 심지어 체크아웃 후에도 시스템에서 데이터를 삭제할 수 있다는 점을 명확하게 설명합니다.

- 사용자 인터뷰 중에 디자인 연구원은 투숙객이 체류하는 동안 호텔 측이 습득한 정보에 대한 요약을 해당 투숙객에게 제공해야 된다는 사실을 알게 됩니다. 투숙객은 체크아웃 시 원하는 경우 호텔에 해당 정보를 삭제하도록 지시할 수 있습니다.

맺음말

AI 디자이너와 개발자는 이러한 5가지 주요 윤리적 영역을 실천함으로써 편향과 권리 박탈을 최소화할 수 있습니다. AI 시스템의 윤리적 문제를 탐지하고 발견되면 수정해야 하므로 지속적인 유지 관리와 개선을 수행할 수 있을 정도로 시스템이 유연해야 합니다.

이 문서에서 다룬 5가지 주요 영역을 도입하고 실천함으로써 디자이너와 개발자는 윤리적인 인식을 높이고 AI 시스템 내 편향을 완화하며 AI를 다루는 사람들에게 책임 의식을 고취할 수 있습니다.

인공지능과 관련하여 우리가 하는 일의 대부분은 우리 모두에게 새로운 영역이므로 개인 및 그룹은 평가를 위한 기준과 지표를 계속 정의하여 문제 탐지와 완화를 지원해야 합니다.

이것은 계속 진행되는 프로젝트입니다. **시간이 지나면서 이 가이드가 발전하고 성숙할 수 있도록 피드백을 보내주시기 바랍니다.** 우리는 이 가이드가 인류를 위한 기술의 의미에 대한 대화와 토론에 기여하고 디자이너와 개발자가 AI 솔루션에 윤리를 적용하는 데 도움이 되기를 바랍니다.

참조 문헌

1. https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/
2. https://ethicsinaction.ieee.org/
3. theinstitute.ieee.org/resources/products-and-services/new-ieee-courses-on-ethics-and-ai-and-autonomous-systems
4. https://arxiv.org/abs/1808.07261
5. http://www.ohchr.org/EN/pages/home.aspx
6. https://ethicsinaction.ieee.org
7. https://www.fastcodesign.com/90164226/what-developers-really-think-about-ai-and-bias
8. https://insights.stackoverflow.com/survey/2018/
9. https://www.ibm.com/design/research/
10. https://www.ethicscanvas.org
11. https://ethicsinaction.ieee.org
12. https://www.ibm.com/ibm/responsibility/2015/at_a_glance/our_approach.html
13. http://faculty.mtsac.edu/cmcgruder/moraljudgements.html
14. https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html
15. https://www.ibm.com/blogs/policy/dataresponsibility-at-ibm/
16. https://martechtoday.com/guide/gdpr-the-general-data-protection-regulation
17. https://woebot.io
18. https://www.ibm.com/blogs/policy/bias-in-ai/
19. www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy-in-america/
20. https://ec.europa.eu/digital-single-market/en/news/eprivacy-consultations-show-confidentiality-communications-andchallenge-new-technologies-are
21. https://www.eugdpr.org/
22. http://www.ohchr.org/EN/ProfessionalInterest/Pages/InternationalLaw.aspx
23. https://ethicsinaction.ieee.org/