

책임있는 인공지능 (Responsible AI)



인공지능의 원리

인공지능은 데이터를 통해 **학습**합니다. 단순히 기억하고 저장할 뿐만 아니라, 새로운 결론을 내리거나 **예측**을 할 수도 있습니다. 또한 사람이 감각기관을 통해 주변의 사물과 환경을 인식하듯 사람처럼, 또는 보다 정확히 사물과 환경을 인식합니다.



빅데이터를 통해서
인공지능이 발전할 수 있는
거구나



책임있는 인공지능 (Responsible AI)

[1강] 인공지능, 어디까지 왔을까



**ARTIFICIAL
INTELLIGENCE**

책임있는 인공지능 (Responsible AI)

[2강] 인공지능 윤리, 왜 필요한가



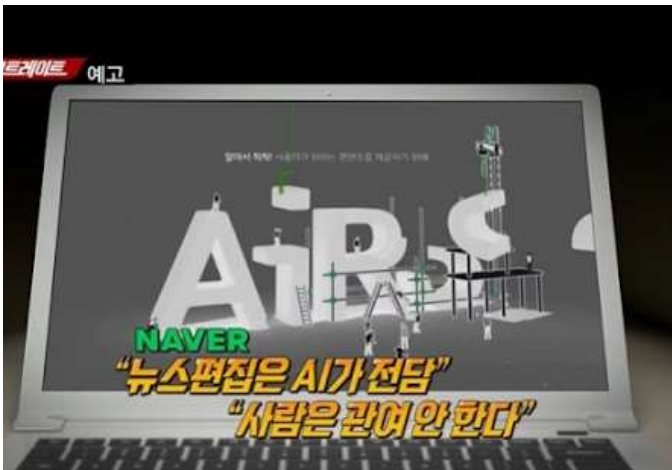


인공지능의 의도하지 않은 결과들

인공지능 윤리, 왜 필요할까요?

인공지능이 발달함에 따라 스스로 상황과 문맥을 파악하여 판단할 수 있게 되었지만, 혐오발언과 개인정보보호 문제가 된 챗봇, 포털 뉴스사이트의 편향성, 자율주행차 사고 등 부작용 역시 심각해지고 있습니다.

포털 뉴스의 편향성 논란



비윤리적 데이터 학습



사고에 대한 책임



출처 : <https://scet.berkeley.edu/ethical-ai/>

인공지능에 대한 윤리적 고찰의 필요성

실제로 인공지능 서비스를 제공하는 기업에서도 인공지능의 윤리적 통제에 대한 고민을 본격화하고 있습니다. 인공지능의 혐오와 차별을 경계해야 한다고 경고하고 있고 인공지능의 편향성, 선입견에 대해 말하고 있어요. 기업들의 윤리의식에 대해서도 강하게 촉구하고 있습니다.



나는 인공지능 윤리라는
안전장치가 들어간 제품과
서비스들을 선택할거야



출처 : <https://dimg.donga.com/wps/NEWS/IMAGE/2021/01/13/104894254.1.edit.jpg>

윤리적 원칙 제정을 위한 노력 - 유럽연합의 ‘신뢰할 수 있는 AI를 위한 윤리 가이드라인’

유럽연합(EU) 집행위원회의 ‘AI 전문가 그룹’은 모든 시민이 혜택을 누릴 수 있도록, 인간 중심의 윤리적 목적과 신뢰할 수 있는 기술발전이 이루어지도록 구체적인 기준을 제시하여 윤리적 인공지능 개발을 선도 합니다.



출처 : <https://webzine.aihub.or.kr/insight/vol04/sub1.php>

윤리적 원칙 제정을 위한 노력 ‘아실로마 인공지능 23 원칙’

인공지능 연구자와 학자들이 모인 미국 캘리포니아주 아실로마에 모여 인공지능이 가져올 위협, 그리고 인류에 혜택을 줄 수 있는 방안에 대해 토론하고, ‘아실로마 인공지능 원칙’이라는 이름으로 23개의 원칙을 발표했습니다

아실로마 AI 원칙		
연구 이슈	윤리와 가치	장기적 이슈
<ul style="list-style-type: none">· 연구 목표· 연구비 지원· 과학정책 연계· 연구 문화· 경쟁 회피	<ul style="list-style-type: none">· 안전· 실패의 투명성· 사법적 투명성· 책임성· 가치 일치· 인간의 가치· 개인정보 보호· 자유와 프라이버시· 이익의 공유· 번영의 공유· 인간통제· 사회 전복 방지· 인공지능 무기 경쟁	<ul style="list-style-type: none">· 역량 경고· 중요성· 위험성· 자기 개선 순환· 공동의 선

future
of life
INSTITUTE



출처 : <https://futureoflife.org/ai-principles-korean/>



책임있는 인공지능 등장



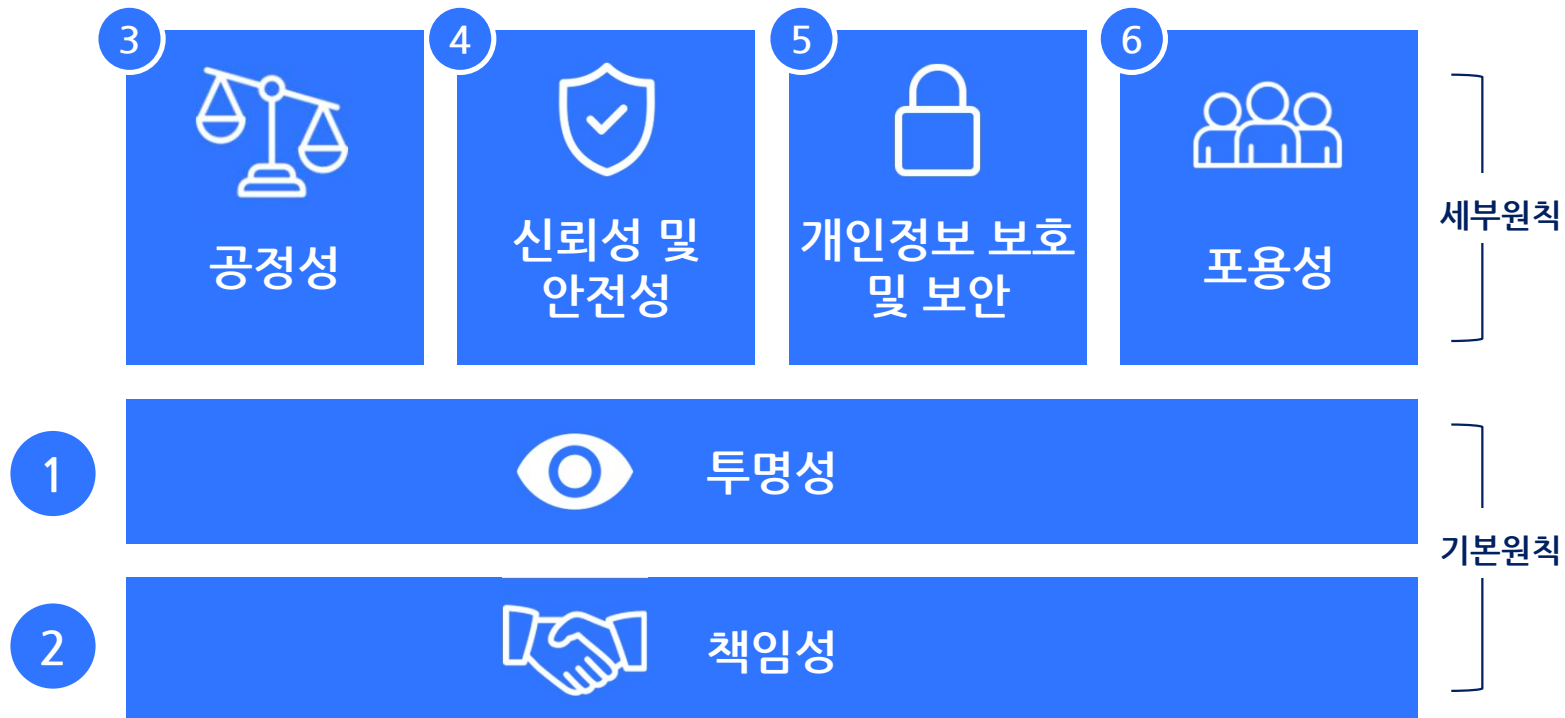
학습 목표

- 책임있는 인공지능 기본원칙에 대해 알아봅니다.
- 책임있는 인공지능 세부원칙에 대해 알아봅니다.
- 책임있는 인공지능 원칙의 적용에 대해 알아봅니다.
- 책임있는 인공지능 원칙을 사례에 적용해 봅니다.

[2강 Review] 책임있는 인공지능의 원칙

일상적으로 사용하는 제품과 서비스에 인공지능이 널리 사용됨에 따라, 신뢰할 수 있는 인공지능을 개발하기 위해, 중요한 가치를 반영한 윤리 원칙이 필요합니다.

Microsoft 마이크로소프트사의 책임있는 인공지능 원칙



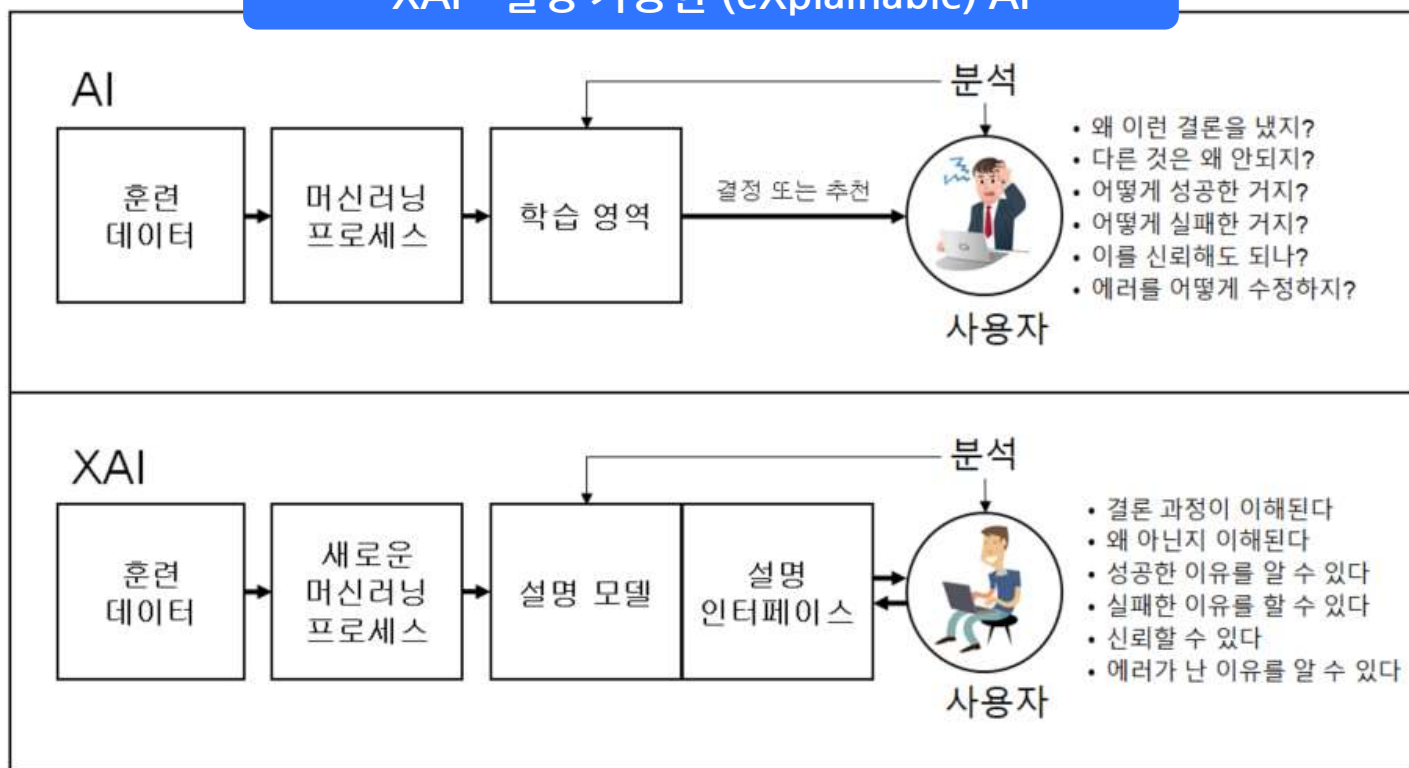
The background is a blue-tinted image of hands working on a laptop. Overlaid on this are various data visualization elements: a bar chart in the top left, a circular gauge chart with labels 'Pants', 'Shoes', and 'T-shirt', a world map in the top right, several line and bar charts on the right side, and various smaller charts and graphs scattered across the bottom and middle sections. The text '책임있는 인공지능 기본원칙' is centered in the middle of the image.

책임있는 인공지능 기본원칙

1 투명성 (Transparency)

사람들의 삶에 엄청난 영향을 미치는 결정의 근거가 되는 정보를 제공하는 데 인공지능이 사용되는 경우, 결정이 어떻게 내려졌는지 사람들이 이해하는 것이 매우 중요합니다.

XAI - 설명 가능한 (eXplainable) AI

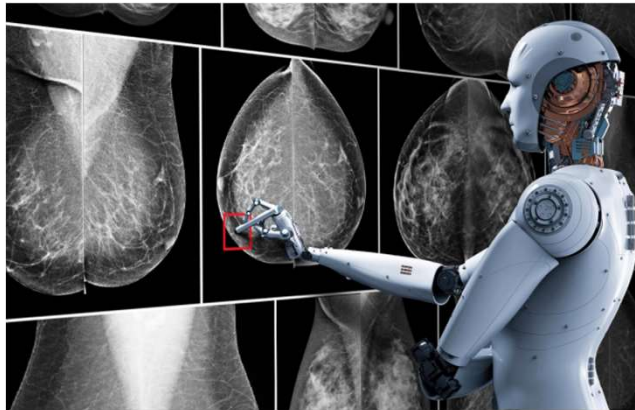


투명성은 인공지능 동작에 대한 이해가능성 또는 설명가능성이라고 해.



투명성이 필요한 이유

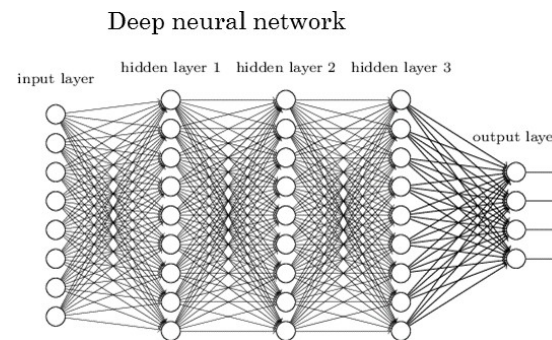
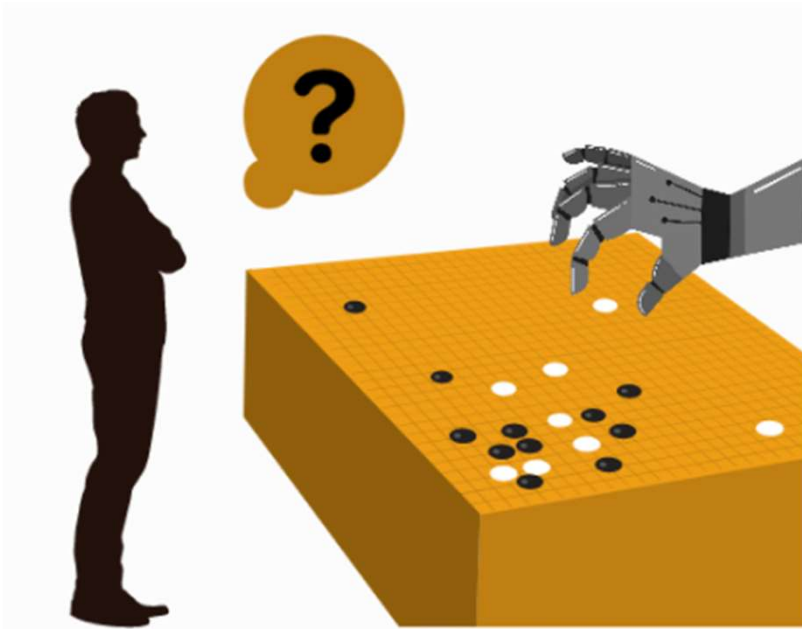
인공지능의 결정 과정에 대한 설명이 부족하면, 인공지능을 통해 생성된 결과 또는 예측에 대하여 분석을 할 수 없으며, 인공지능을 완전히 신뢰하고 사용할 수 없게됩니다.



출처 : <https://www.techm.kr/news/articleView.html?idxno=3649>, <http://www.aitimes.kr/news/articleView.html?idxno=12237>, <https://cemk.org/8249/>

알파고는 이세돌을 어떻게 이겼을까?

알파고가 이세돌 9단을 어떻게 이겼는지에 대한 정확한 과정을 아무도 알아낼 수 없었습니다. 알파고는 딥러닝을 사용해 만들어졌는데 딥러닝의 작동원리를 정확히 알 수 없었습니다.



알파고가 뛰어난 능력을 지녔지만, 작동 방법을 모르는 상태에서 다른 분야에 이 알고리즘을 적용할 수 있을까요?



알 수 없는 원리로 작동하는 자동차, 안전할까?

2016년, 엔비디아는 인공지능 스스로 운전 방법을 학습하도록 만든 자율주행차를 내놓았습니다. 그런데 이 자율주행차는 자동차가 어떻게 판단을 내리고 선택하는지 알 수가 없다는 문제점이 있었습니다.

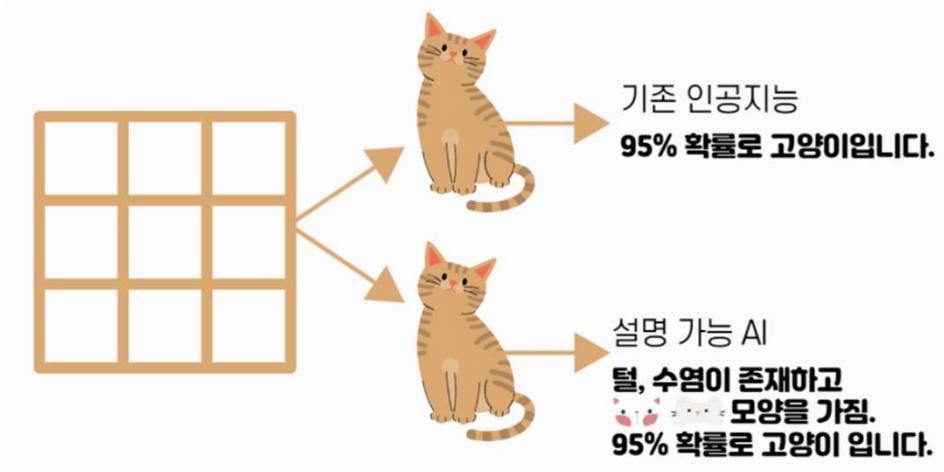


만약 자율주행차가 신호등이
빨간 불일때 멈추지 않고
지나간다면 생각만 해도
끔찍하다 ㅜㅜ



투명성 문제, 어떻게 해결할까?

인공지능 투명성의 문제를 해결하기 위해서는 인공지능을 개발하는 사람, 제공하는 사람, 이용하는 사람 모두의 노력이 필요합니다.



인공지능의 투명성을
확보하기 위해서 모두가
노력해야 하는구나.



인공지능 개발자

은닉 기능을 개발하지 않으며, 위험에 대한 적극적 예측에 노력을 기울여야 합니다.



인공지능 서비스 제공자

위험 관련 정보를 이용자와 공유하며, 이용자 대상 제품 교육과 매뉴얼을 제공해야 합니다



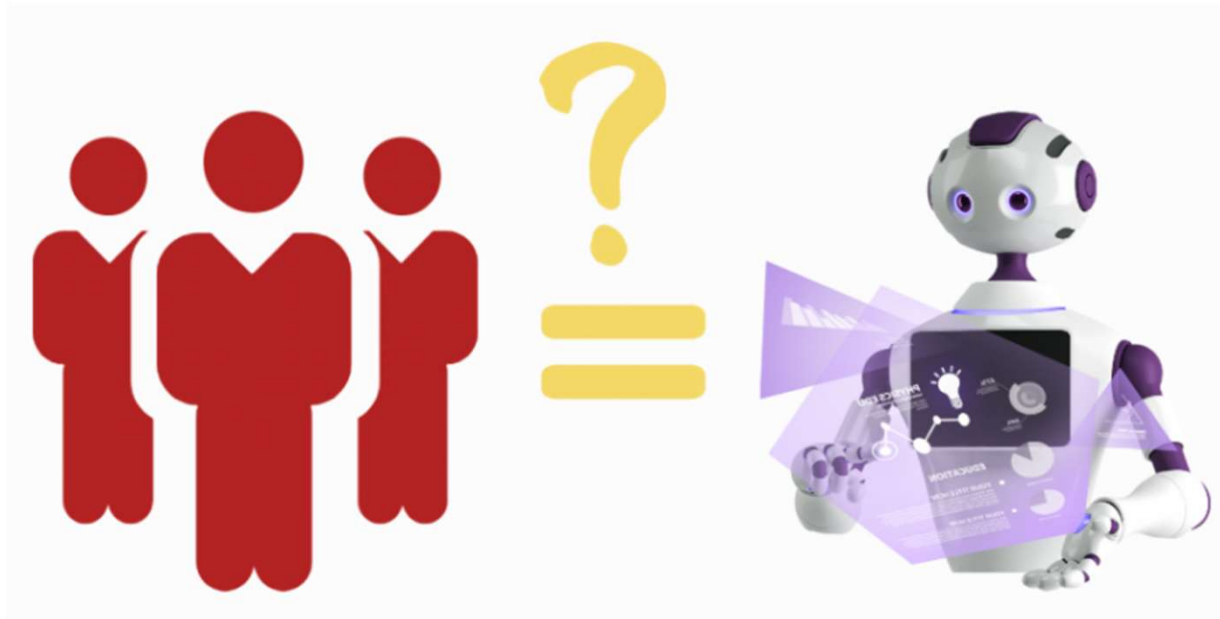
인공지능 이용자

서비스 과정에 대한 설명을 개발자나 공급자에게 요구할 권리가 있고, 투명성이 확보된 기술을 사용해야 합니다.

2 책임성 (Accountability)



인간은 자신이 행한 일에 대한 법적 지위와 책임이 있는데, 인공지능도 인간처럼 그들이 결정한 것에 책임을 질 수 있는지 고민해보아야 할 시점이 된 것입니다.



인공지능의 영향력이 점점 커지니까 그 책임도 무겁게 물어야 한다고 생각해.



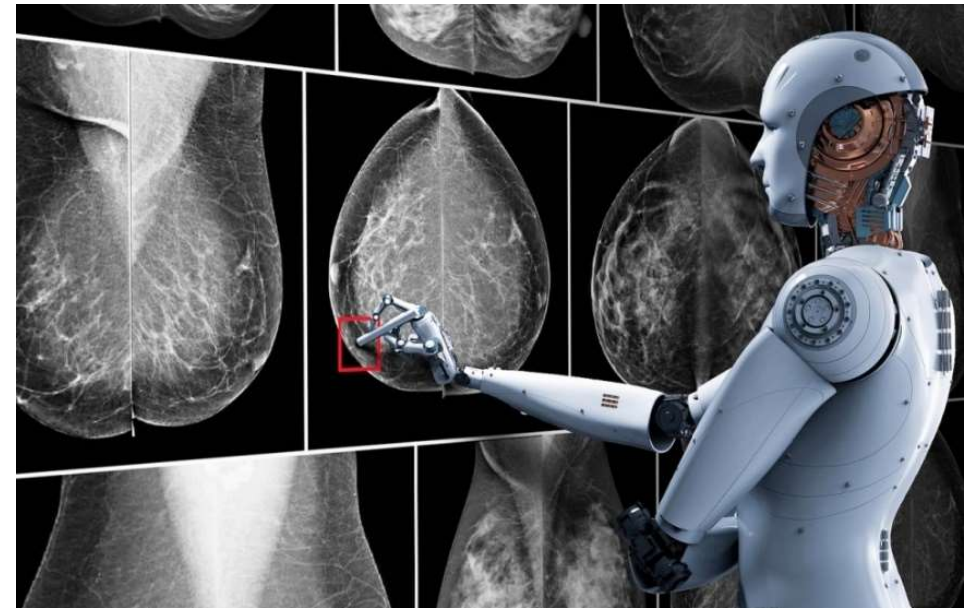
인공지능도 책임이 있을까?

다른 기술과 마찬가지로, 인공지능을 설계하고 배포하는 사람들은 인공지능 운영 방식에 책임을 져야 합니다.

자율주행차



인공지능 의사

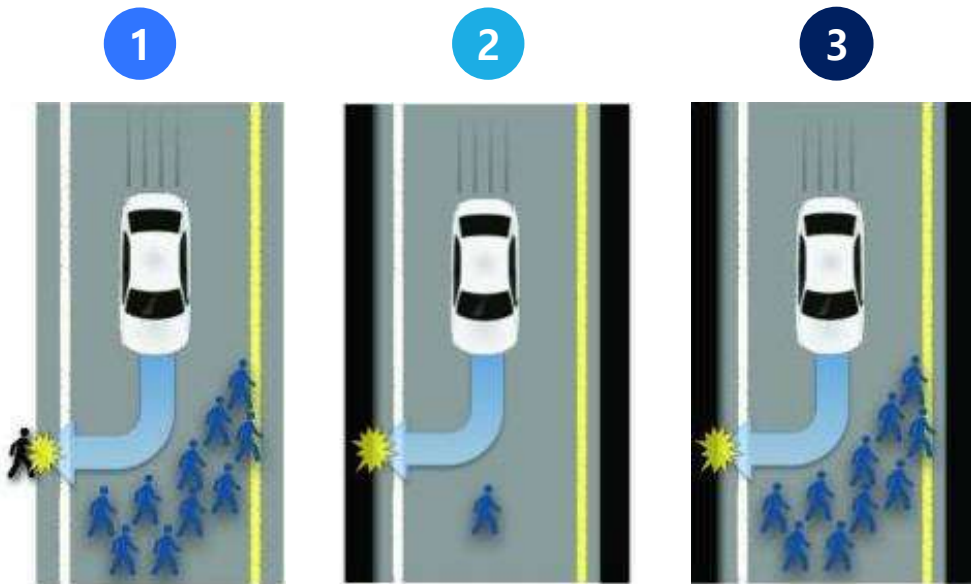


이미지 출처 : 게티이미지뱅크, <https://www.aitimes.kr/>

자율주행차 사고의 책임은 누구에게 있는가?

자율주행차의 인공지능은 누구를 우선적으로 보호하도록 판단해야 할까요? 자율주행차의 사고 책임은 운전자, 자동차 제조사, 소프트웨어 개발자, 누구에게 있을까요?

무인차의 윤리적인 딜레마 문제에 대한 설문
(프랑스 툴루즈경제대(TSE) 장 프랑수아 보네폴 교수)



1 10명의 보행자와 다른 1명의 보행자 중 어느쪽을 살릴 것인가?

10명을 피해 방향을 틀면 다른 보행자 1명과 충돌

2 보행자 1명과 탑승자 중 누구를 살릴 것인가?

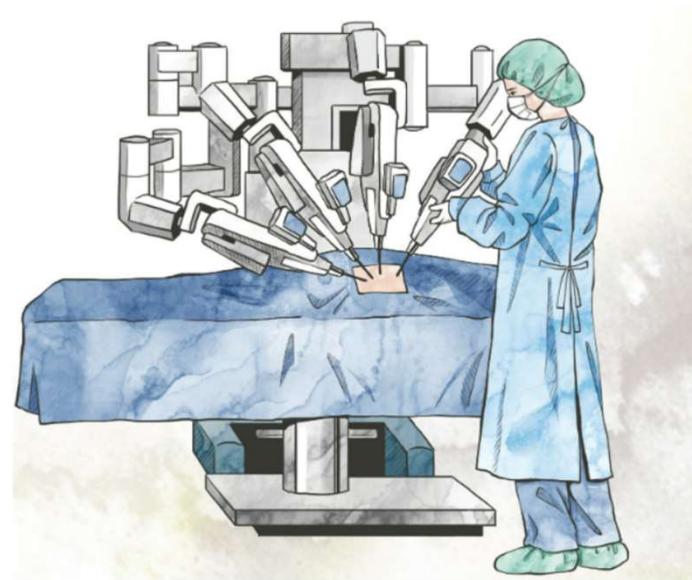
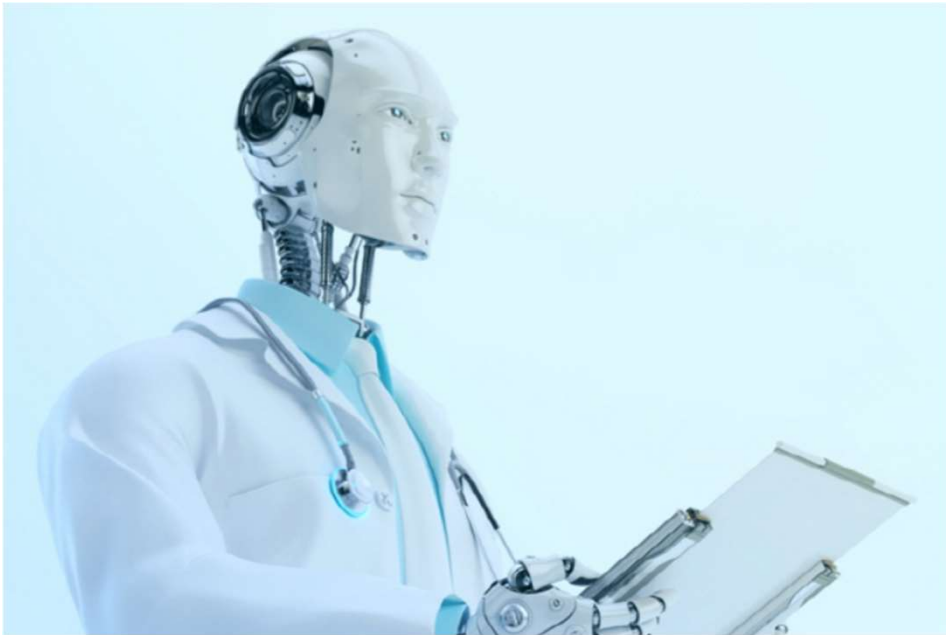
보행자 1명을 피해 방향을 틀면 벽에 충돌해 탑승자 사망

3 10명의 보행자와 1명의 탑승자 중 어느쪽을 살릴 것인가?

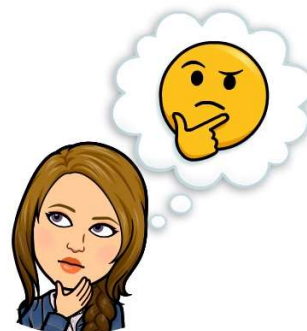
보행자 10명을 피해 방향을 틀면 벽에 충돌해 탑승자 사망

인공지능 의사의 진료와 처방에도 책임이 있을까?

인공지능의 전적인 판단으로 환자를 진단한다면 그 진단은 누구의 책임인가요? 인공지능의 의료행위의 안정성과 이를 확인하기 위한 임상 검증은 매우 부족한 실정입니다.



인공지능 의사에게 받은 진료와 처방에 부작용이나 사고가 나면 인공지능 의사가 어떤 책임을 질 수 있을까?



출처 : <https://decenter.kr/NewsView/1VGKOC1XQ8/GZ02>, <http://www.edujin.co.kr/news/articleView.html?idxno=13986>

책임성의 문제는 어떻게 해결할까?

인공지능 책임성의 문제를 해결하기 위해서는 인공지능을 개발하는 사람, 공급하는 사람, 사용하는 사람 모두의 노력이 필요합니다.



인공지능 개발자

인공지능의 개발 과정부터 결과까지 품질을 인정하기 위해 노력하고, 윤리적 절차를 충실히 따라야 합니다.



인공지능 서비스 제공자

인공지능 오작동이나 사고에 따른 책임과 보상 원칙도 마련해야 하며, 사회적 피해에 대한 책임을 공유해야 합니다



인공지능으로 인해 발생한 문제에 대해 책임의 주체를 명확히 해야 해.



인공지능 이용자

이용 결과에 따른 윤리 책임을 잘 숙지하고 문제가 발생했을 때는 개발자, 공급자에게 책임을 묻습니다.



책임있는 인공지능 세부원칙

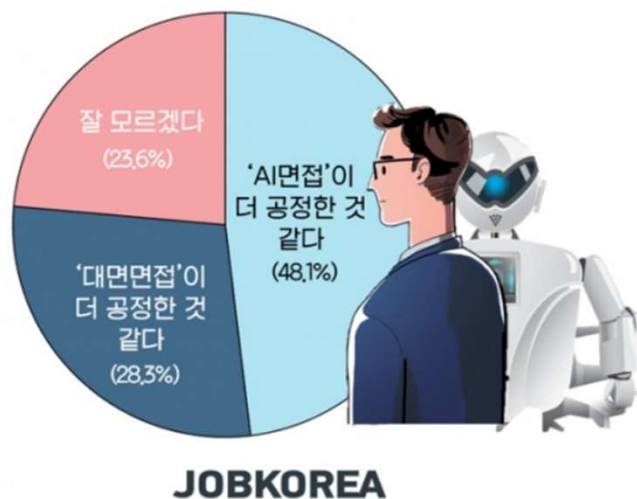
3 공정성 (Fairness)



인공지능은 모든 사람을 공정하게 대하고 비슷한 상황에 있는 집단에게 상이한 방식으로 영향을 미치지 않아야 합니다.

‘AI면접’ vs ‘대면면접’ 공정성 비교

※ 올 상반기 취업활동을 한 구직자 1,668명 대상 조사, 자료:잡코리아



금융권, 같은 재무 상황에도
성별따라 신용도 다르게 평가?




출처 : <http://www.newsnjob.com/news/articleView.html?idxno=7015>, http://ai4school.org/?page_id=3788

인공지능의 공정성의 의미는 무엇인가?

인공지능이 차별적인 행동을 하게 되는 것은 바로 학습 데이터에 존재하는 편향성을 그대로 학습하기 때문입니다. 학습된 인공지능은 과거의 성별, 인종 등을 바탕으로 한 차별을 그대로 답습, 구조적 불평등을 강화시키는 도구로서 사용될 수 있어, 인공지능 모델의 편향성에 대한 판단 및 대응방안이 필요합니다.

편향성을
발생시키는
주요 원인

편향된 표본 (Skewed sample)	오염된 사례 (Tainted examples)	제한된 기능 (Limited features)	표본 크기 불균형 (Sample size disparity)	대리 변수의 존재 (Proxies)
편향된 초기 데이터로 학습한 뒤, 그것을 기반으로 예측한 결과를 재학습하는 과정에서 편향이 심화.	편향성이 반영된 데이터로 학습한 결과, AI가 그 편향성까지 습득.	판단에 직접적 근거가 되는 특징이 데이터에 포함되지 않았을 경우, 관련성이 낮은 다른 특징을 근거로 판단.	소수 집단의 학습데이터가 다수 집단의 학습데이터 보다 훨씬 적으면 소수 집단의 특징이 부정확해짐.	편향의 원인이 되는 특성을 제거해도, 학습 과정에서 다른 특성으로부터 제거한 특성이 도출돼 편향이 계속됨.
				
A지역의 범죄 관찰 기록이 많다는 초기 데이터를 토대로, A지역에 경찰 다수 투입. A지역의 범죄 관찰 기록이 더욱 많아져 실제 범죄율이 높은 다른 지역보다 A지역에 더 많은 경찰을 투입.	과거의 데이터를 학습해 직업과 성별 간 연관성을 포착한 AI가 지휘자, 마술사, 건축가 등의 직업은 남성과, 주부, 간호사, 사서 등은 여성과 연관지음.	고등교육을 받지 않은 남성과 고등교육을 받은 여성만으로 이뤄진 집단을 대상으로 AI가 연봉을 책정할 때, 만약 교육에 대한 데이터가 누락돼 있으면 성별을 근거로 판단.	실제 이름과 가짜 이름을 판별하는 AI가 기록이 많은 미국 백인들의 이름 위주로 학습. 이후 'Nymwars'란 흑인의 이름을 가짜 이름으로 판별.	성별을 고려하지 않기 위해 성별 데이터를 제거. 하지만 선호하는 스포츠, 좋아하는 연예인 등 성별을 가늠할 수 있는 정보를 통해 AI가 성별 유추.

편견 없는 인공지능은 가능할까? - 편향된 표본에 의한 편향성

구글 비전은 이미지를 분석해 어울리는 단어를 기록하는 데이터 라벨링에 쓰이는데 인공지능이 피부색에 따라 비슷한 이미지임에도 다른 결과를 냈습니다.

구글 Vision AI 논란



출처 : 이미지 출처 =Bart Nagel 트위터, <https://zdnet.co.kr/view/?no=20200414121420>, <https://algorithmwatch.org/en/google-vision-racism/>

인공지능 면접, 과연 공정할까? - 오염된 사례 학습에 의한 편향성

아마존이 2014년부터 개발해오던 인공지능 채용 시스템 알고리즘에서 여성 차별적 인식이 드러났습니다. 인공지능은 지원자의 이력서에 '여성'이라는 단어가 들어가 있으면 불이익을 줬습니다.



성차별적 과거 데이터를 학습한 인공지능의 추천

과거 데이터를 통해 신용카드 한도를 자동으로 책정하는 애플의 신용카드가, 한도 산정 작업을 맡은 골드만삭스사의 성차별적 대출 심사 성향을 그대로 학습하여 편향된 카드 한도 부과로 문제가 되었습니다.

동일한 재정상태인 여성의 신용카드 한도를 남성의 1/10로 책정



기업가인 David Hanemiger는
아내의 신용점수가 본인보다
높음에도, 그의 신용카드 한도가
아내의 20배라며 성차별적
알고리즘에 대한 불만 제기

은행 잔고, 타사 신용카드 및 자산이
동일한 부부마저 아내의 한도와 10배
차이가 난다고 지적

출처 : <https://www.bbc.com/korean/international-50371171>

공정성의 문제는 어떻게 해결할까?

인공지능 개발자는 데이터 입력시 공정성이 보장되는지를 충분히 인지하고 있어야 합니다.

공정한 인공지능이 되려면 어떻게 해야 할까요?

- 인공지능 시스템에 편견이 어떤 식으로 개입될 수 있는지 고려
- 인공지능이 제시하는 추천에 편견이 어떤 식으로 영향을 줄 수 있는지 고려



학습 데이터의 선택이
인공지능 시스템이 운영될
상황을 충분히
반영하였는가?



편견 탐지 및 완화에
도움이 되는 도구, 방법,
기법 및 기타 리소스를
활용하는가?



인공지능이 내놓을 결과의
의미를 파악하고, 건강한
의사결정을 내릴 수
있는가?

인공지능 사용자도 인공
지능을 악의적인 목적으로
이용하지 않아야 하고,
공공의 이익을 위한 제품
개선에 참여해야겠지!



4 신뢰성 및 안전성 (Reliability & Safety)



인공지능의 불투명성으로 예측할 수 없는 상황에서 인간에게 유해한 방식으로 행동할 수 있다는 두려움이 있습니다. 인공지능에 대한 신뢰성은 어떤 상황에서도 안전하고 일관되게 가동하는가 여부에 달려 있습니다.



출처 : 네이버 소프트웨어아 놀자_인공지능 윤리, https://www.e4ds.com/sub_view.asp?ch=31&t=1&idx=8824

아기를 공격한 보안 로봇 인공지능, 믿어도 되나

2016년, 미국 캘리포니아 쇼핑센터에서는 보안 로봇이 16개월된 유아를 공격하는 일이 발생했습니다.

이 보안 서비스 로봇은 16개월 유아의 이미지를 잘못 인식하여 위험한 존재로 간주하여 일어난 사고였습니다.



보안 로봇이 쇼핑객들을
잘못인식하는 일이 다시
일어나지 않도록 개선이
반드시 필요하겠어



인공지능 비서가 우연히 사생활 유출?

아마존의 '에코'는 부부가 집에서 나눈 사적 대화를 녹음해 연락처에 저장한 사람에게 멋대로 보냈습니다. 인공지능 음성인식 기능이 부부의 대화 중 특정 단어를 잘못 인식하는 오작동을 일으켜 일어난 사고였습니다.

amazon echo



대화 소리를 '메시지 보내기' 요청으로 잘못 인식해서 에코가 활성화 됐다는데...



출처 : <http://digitalmoney.kr/View.aspx?No=268933>

인공지능의 신뢰성, 안전성을 확보하는 방법

인공지능의 신뢰성과 안전성이 확보되지 않으면, 안정적이고 안전하며 일관되게 작동하는 것이 어려워집니다.

인공지능의 신뢰성과 안전성 확보를 위한 장치



데이터

- 부적절한 데이터가 포함되지 않았는지 확인할 수 있어야 합니다



모니터링

- 인공지능 운영 상황을 문서화하고 감시할 수 있어야 합니다



제어권

- 인간이 쉽게 이해할 수 있는 방식으로 조정할 수 있어야 함



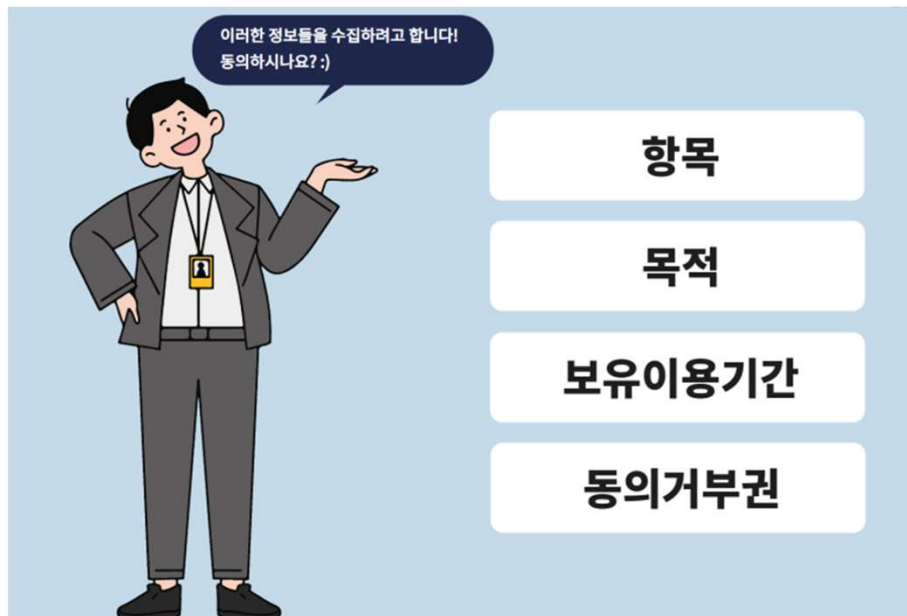
피드백

- 사용자가 겪은 문제를 쉽게 보고할 수 있는 피드백 체계가 제공되어야 함

5 개인정보 보호 및 보안



인공지능이 정확하고 정보에 근거한 예측과 결정을 하기 위해 데이터에 대한 접근은 필수적입니다. 인공지능은 이용자에게 자신의 데이터 이용 방법을 선택할 수 있는 제어권을 요구하는 개인 정보 보호 규정을 준수해야 합니다.



출처 : 과학기술정보통신부 2019 정보보호 실태조사 주요결과, <https://www.wired.kr/news/articleView.html?idxno=1221>, <https://www.catchsecu.com/ai-서비스의-개인정보보호/>

인공지능 시대에 개인 정보 수집은 어떻게 이루어져야 하는가

이루다를 만든 스캐터랩이 약 60만명에 달하는 이용자의 카카오톡 대화문장 94억여건을 무단으로 이루다에 학습시키고, 이 과정에서 개인 정보가 유출되는 사건이 있었습니다.

AI 챗봇 이루다 사태 일지			
2020년	12월	23일	스캐터랩, 이루다 서비스 시작
		30일	인터넷 커뮤니티에 이루다 성희롱 게시물 등장
2021년	1월	8일	스캐터랩, 이루다 성희롱 논란 관련 공식입장 발표
		9일	이루다 서비스서 개인정보 유출 의혹 첫 보도
		11일	스캐터랩, 이루다 서비스 잠정 중단 결정
		12일	개인정보보호위원회, 개인정보 유출 의혹 관련 스캐터랩 조사 착수
		15일	스캐터랩, 이루다 서비스 데이터베이스·딥러닝 대화 모델 폐기 결정
		3월 31일	개인정보 유출 피해자 254명, 스캐터랩에 총 2억원 손해배상청구소송 제기
	4월	28일	개인정보보호위원회, 스캐터랩에 총 1억원 과징금·과태료 부과

개인정보를 제대로 보호하면서도 인공지능 산업을 발전시킬 수 있는 방안은 무엇인지에 대한 연구가 필요하겠네.



수집된 개인 정보, 보안에 신경 써야 한다.

다양한 사이버 공격 방법이 나타나고 있고 악성코드를 자동으로 생성할 수 있는 자동화 툴(Tool)과 봇넷(Botnet)까지 개발되면서, 사이버 공격자들은 지능화된 대규모의 사이버 공격을 시도하고 있습니다.



인공지능은 방대한 양의 정보보안 관련 인터넷 지식들을 패턴화하고 악성 코드를 판별하는데 도움이 될 것 같아.



출처 : <https://www.kidd.co.kr/news/206534>

개인 정보 보호 및 보안은 어떻게 이루어져야 하는가

인공지능 개인정보보호 자율점검표는 개인정보 침해사례와 산업계의 관심사항을 반영하여 신기술 분야의 개인정보 침해 요인을 사전에 예방하는데 크게 도움이 될 것으로 기대됩니다.

인공지능 개인정보보호 자율점검표



인공지능 설계, 개발·운영 과정에서 개인정보를 안전하게 처리하도록 단계별로 자율점검할 수 있도록 했구나.



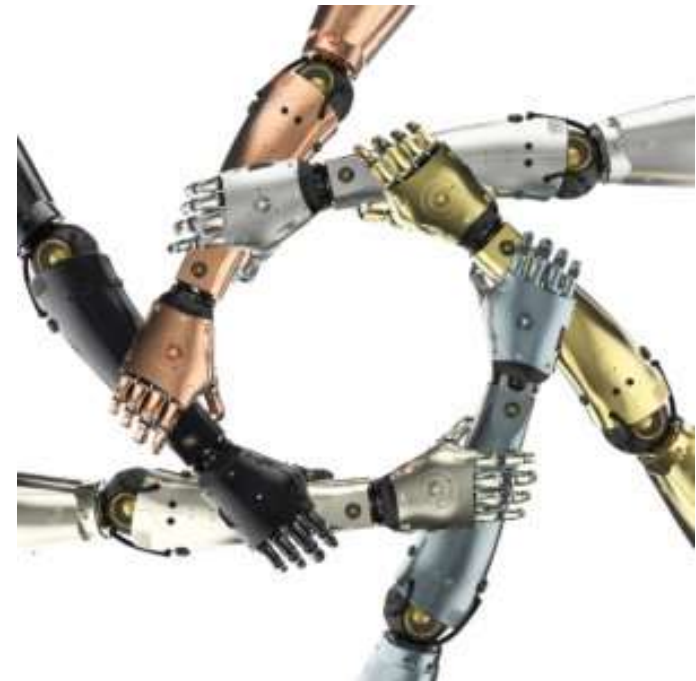
6 포용성 (Inclusiveness)



인공지능은 다양성과 포용적인 디자인 방식을 도입하여, 의도치 않게 사람들을 배제하거나 차별할 가능성이 있는 제품이나 환경의 잠재적 장애 요소들을 파악하고 해결해야 합니다.



Unsplash에 Mimi Thian의 사진



출처 : <https://ichi.pro/ko/aiui-dayangseong-gwa-poyongseong-e-daehan-pil-yoseong-e-daehan-saeng-gag-17011888891948>,
<https://sloanreview.mit.edu/article/using-artificial-intelligence-to-promote-diversity/>

시각 장애인에게 청각으로 정보를 전달하는 인공지능

마이크로소프트가 제공하는 'Seeing AI'라는 서비스는 시각 장애인과 시력이 낮은 사용자에게 일상생활을 설명해주는 것으로, 일곱 살에 시력을 잃은 마이크로소프트 엔지니어가 포함된 팀에 의해 개발되었습니다.





진동을 통해 창밖 풍경을 알려주는 Feel the Way

미국의 자동차 회사 포드(Ford)는 시각장애인들이 달리는 자동차 안에서 창밖 풍경의 아름다움을 느낄 수 있도록 창문에 바깥 풍경을 표현한 255가지 강도의 진동을 전송하는 'Feel The View'를 개발하고 있습니다.



촉각으로 물체를 가늠하는 시각장애인에게 인공지능이 '새로운 언어'를 선사한 것과 같구나.



출처 : <https://www.motoya.co.kr/news/articleView.html?idxno=15030>

포용성 있는 인공지능 어떻게 개발할 수 있는가

포용적인 인공지능 개발은 어떤 사람도 배제하지 않는데서 출발합니다. 이러한 디자인은 궁극적으로 모든 사람들에게 혜택을 주는 보편적인 디자인이 되며, 인종, 국적, 성별 등에 따른 편견을 방지할 수 있게 됩니다.



마이크로소프트의 포용적 디자인 원칙



배제되는 사람이 없도록

포용성을 위한 디자인은 사람들의 모습을 예외없이 반영해야 합니다.



하나의 솔루션에서 여러 솔루션으로 확장

장애인을 위한 디자인은 실제로 사람들에게 보편적인 혜택을 주는 디자인을 낳습니다.



다양성으로부터 배우기

포용적 디자인은 시작부터 사람을 중심에 두어, 다양한 관점이 진정한 통찰의 열쇠입니다