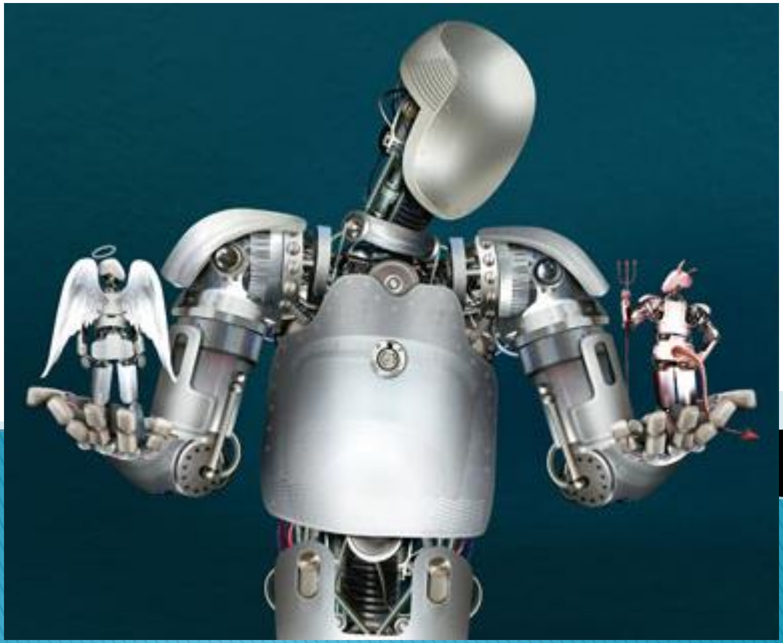
Institute of
Social Computing
소셜컴퓨팅연구소

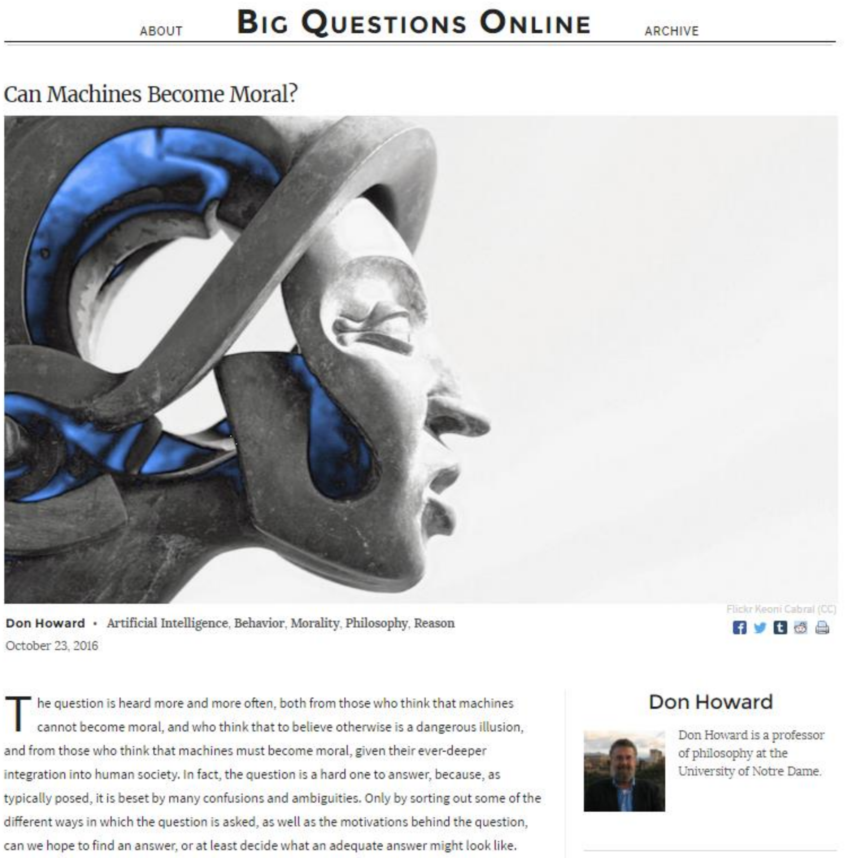
인공지능은 도덕적 기계가 될 수 있는가?

2017. 4



[이미지 출처:
The Economist]

AI는 Moral Machine일 수 있는가?



ABOUT **Big Questions Online** ARCHIVE

Can Machines Become Moral?

Flickr Keoni Cabral (CC)

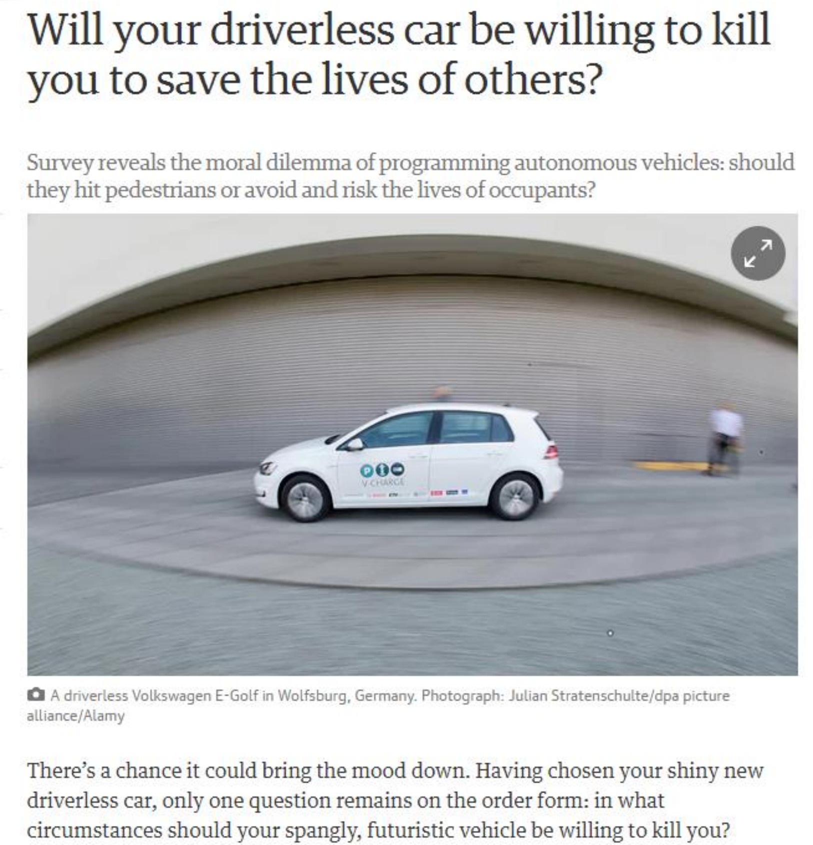
Don Howard • Artificial Intelligence, Behavior, Morality, Philosophy, Reason

October 23, 2016

The question is heard more and more often, both from those who think that machines cannot become moral, and who think that to believe otherwise is a dangerous illusion, and from those who think that machines must become moral, given their ever-deeper integration into human society. In fact, the question is a hard one to answer, because, as typically posed, it is beset by many confusions and ambiguities. Only by sorting out some of the different ways in which the question is asked, as well as the motivations behind the question, can we hope to find an answer, or at least decide what an adequate answer might look like.

Don Howard

Don Howard is a professor of philosophy at the University of Notre Dame.



Will your driverless car be willing to kill you to save the lives of others?

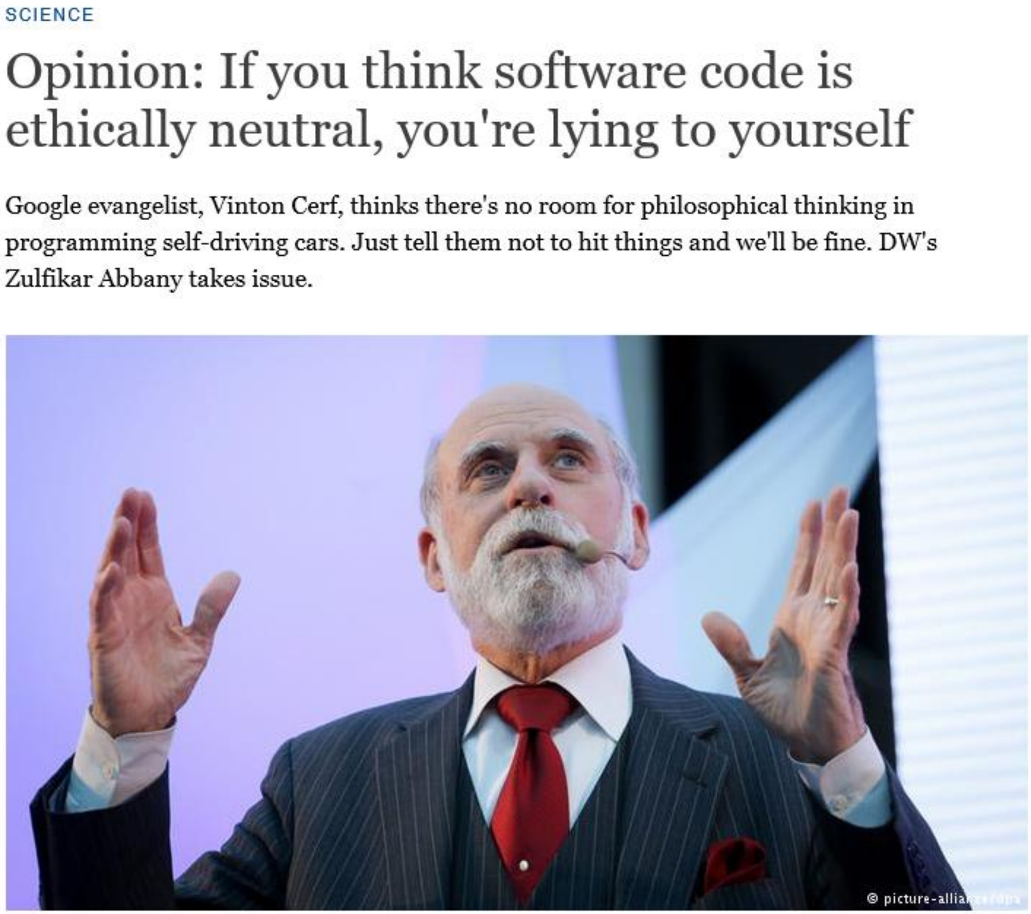
Survey reveals the moral dilemma of programming autonomous vehicles: should they hit pedestrians or avoid and risk the lives of occupants?

A driverless Volkswagen E-Golf in Wolfsburg, Germany. Photograph: Julian Stratenschulte/dpa picture alliance/Alamy

There's a chance it could bring the mood down. Having chosen your shiny new driverless car, only one question remains on the order form: in what circumstances should your spangly, futuristic vehicle be willing to kill you?

[출처: The Guardian]

The Software Code Is No Ethically Neutral



SCIENCE

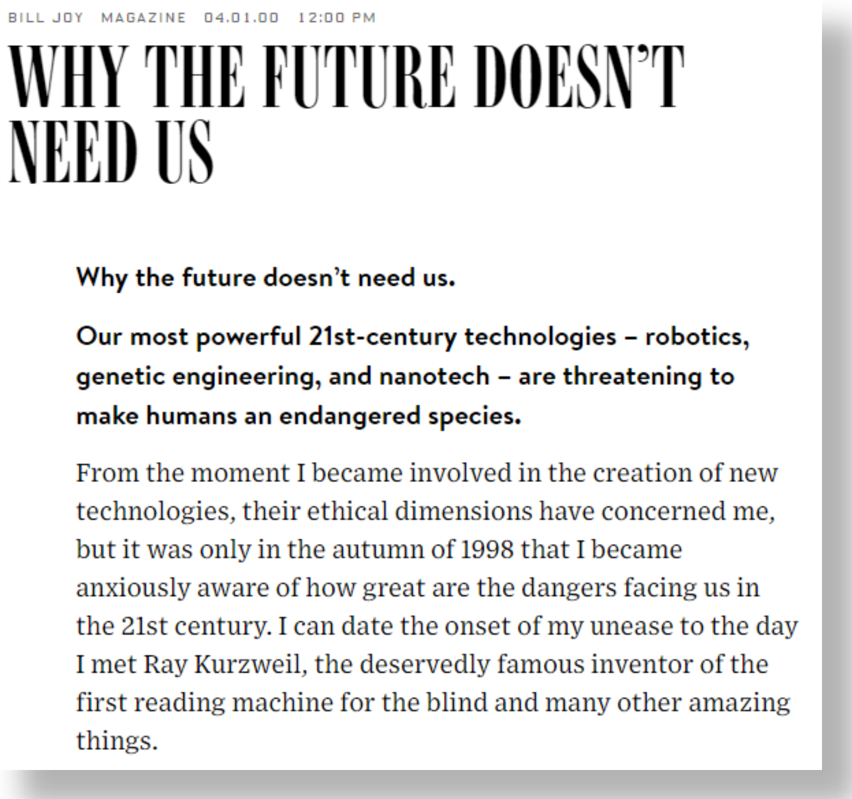
Opinion: If you think software code is ethically neutral, you're lying to yourself

Google evangelist, Vinton Cerf, thinks there's no room for philosophical thinking in programming self-driving cars. Just tell them not to hit things and we'll be fine. DW's Zulfikar Abbany takes issue.

© picture-alliance/dpa

Bill Joy’s Warning in 2000

- “The only realistic alternative I see is relinquishment: to limit development of the technologies that are too dangerous, by limiting our pursuit of certain kinds of knowledge”



BILL JOY MAGAZINE 04.01.00 12:00 PM

WHY THE FUTURE DOESN’T NEED US

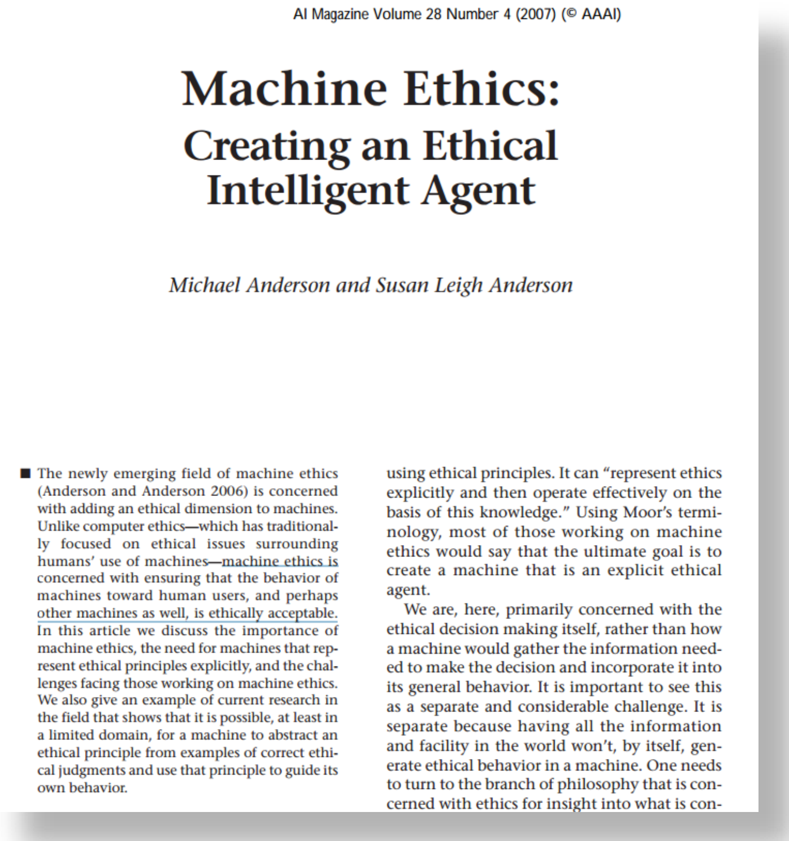
Why the future doesn’t need us.

Our most powerful 21st-century technologies – robotics, genetic engineering, and nanotech – are threatening to make humans an endangered species.

From the moment I became involved in the creation of new technologies, their ethical dimensions have concerned me, but it was only in the autumn of 1998 that I became anxiously aware of how great are the dangers facing us in the 21st century. I can date the onset of my unease to the day I met Ray Kurzweil, the deservedly famous inventor of the first reading machine for the blind and many other amazing things.

Machine Ethics

- Machine ethics is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable
- The ultimate goal of machine ethics is to create a machine that itself follows an ideal ethical principle or set of principles; that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of action it could take
- Implicit vs. Explicit (Moor, J. H. 2006. The Nature, Importance, and Difficulty of Machine Ethics. IEEE Intelligent Systems 21(4): 18–21.)
  - A machine that is *an implicit ethical agent* is one that has been programmed to behave ethically, or at least avoid unethical behavior, without an explicit representation of ethical principles -- constrained in its behavior by its designer who is following ethical principles
  - A machine that is *an explicit ethical agent* is able to calculate the best action in ethical dilemmas using ethical principles
- Machine ethics is an inherently interdisciplinary field



AI Magazine Volume 28 Number 4 (2007) (© AAAI)

Machine Ethics:
Creating an Ethical Intelligent Agent

Michael Anderson and Susan Leigh Anderson

■ The newly emerging field of machine ethics (Anderson and Anderson 2006) is concerned with adding an ethical dimension to machines. Unlike computer ethics—which has traditionally focused on ethical issues surrounding humans' use of machines—machine ethics is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable. In this article we discuss the importance of machine ethics, the need for machines that represent ethical principles explicitly, and the challenges facing those working on machine ethics. We also give an example of current research in the field that shows that it is possible, at least in a limited domain, for a machine to abstract an ethical principle from examples of correct ethical judgments and use that principle to guide its own behavior.

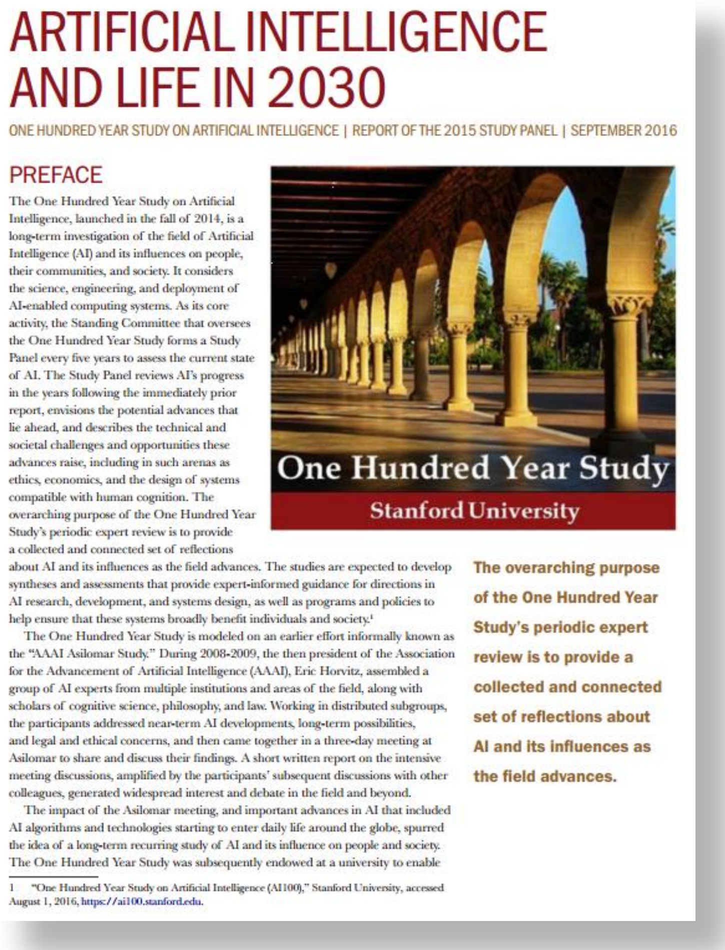
using ethical principles. It can "represent ethics explicitly and then operate effectively on the basis of this knowledge." Using Moor's terminology, most of those working on machine ethics would say that the ultimate goal is to create a machine that is an explicit ethical agent.

We are, here, primarily concerned with the ethical decision making itself, rather than how a machine would gather the information needed to make the decision and incorporate it into its general behavior. It is important to see this as a separate and considerable challenge. It is separate because having all the information and facility in the world won't, by itself, generate ethical behavior in a machine. One needs to turn to the branch of philosophy that is concerned with ethics for insight into what is con-

The Importance of Machine Ethics

- Ethical ramifications to what machines currently do and are projected to do in the future
- Humans' fear – whether these machines will behave ethically, so the future of AI may be at stake
- Research in machine ethics will advance the study of ethical theory
  - AI makes philosophy honest – Daniel Dennett
  - How agents ought to behave in ethical dilemmas
- "Because we are concerned with machine behavior, we can be more objective in examining ethics than we would be in discussing human behavior" – Susan Leigh Anderson, Professor Emerita of Philosophy at the University of Connecticut

AI의 사회적 영향에 대한 본격적 논의



ARTIFICIAL INTELLIGENCE AND LIFE IN 2030

ONE HUNDRED YEAR STUDY ON ARTIFICIAL INTELLIGENCE | REPORT OF THE 2015 STUDY PANEL | SEPTEMBER 2016

PREFACE

One Hundred Year Study

Stanford University

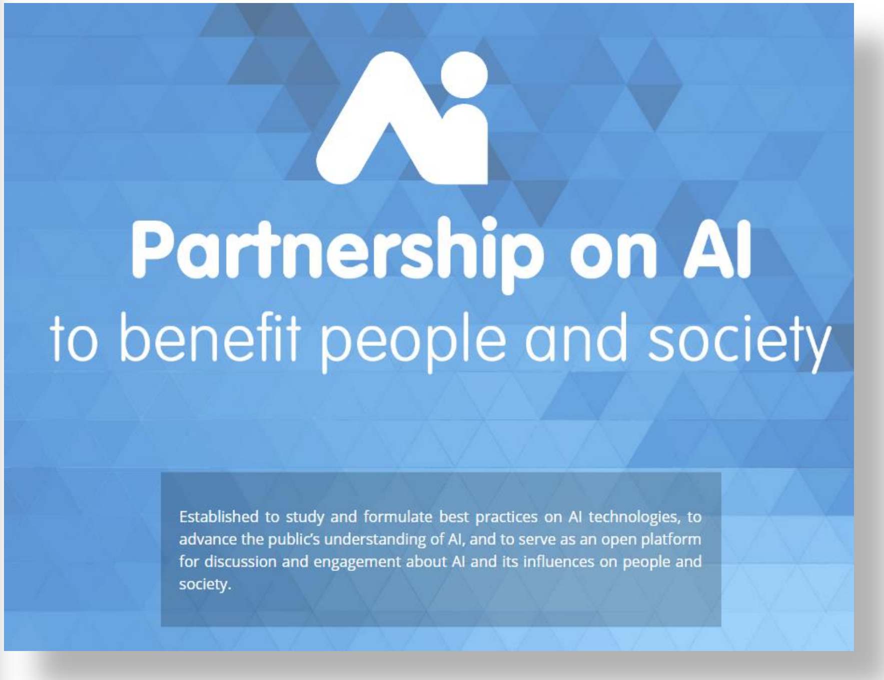
The One Hundred Year Study on Artificial Intelligence, launched in the fall of 2014, is a long-term investigation of the field of Artificial Intelligence (AI) and its influences on people, their communities, and society. It considers the science, engineering, and deployment of AI-enabled computing systems. As its core activity, the Standing Committee that oversees the One Hundred Year Study forms a Study Panel every five years to assess the current state of AI. The Study Panel reviews AI's progress in the years following the immediately prior report, envisions the potential advances that lie ahead, and describes the technical and societal challenges and opportunities these advances raise, including in such arenas as ethics, economics, and the design of systems compatible with human cognition. The overarching purpose of the One Hundred Year Study's periodic expert review is to provide a collected and connected set of reflections about AI and its influences as the field advances. The studies are expected to develop syntheses and assessments that provide expert-informed guidance for directions in AI research, development, and systems design, as well as programs and policies to help ensure that these systems broadly benefit individuals and society.[1]

The One Hundred Year Study is modeled on an earlier effort informally known as the "AAAI Asilomar Study." During 2008-2009, the then president of the Association for the Advancement of Artificial Intelligence (AAAI), Eric Horvitz, assembled a group of AI experts from multiple institutions and areas of the field, along with scholars of cognitive science, philosophy, and law. Working in distributed subgroups, the participants addressed near-term AI developments, long-term possibilities, and legal and ethical concerns, and then came together in a three-day meeting at Asilomar to share and discuss their findings. A short written report on the intensive meeting discussions, amplified by the participants' subsequent discussions with other colleagues, generated widespread interest and debate in the field and beyond.

The impact of the Asilomar meeting, and important advances in AI that included AI algorithms and technologies starting to enter daily life around the globe, spurred the idea of a long-term recurring study of AI and its influence on people and society. The One Hundred Year Study was subsequently endowed at a university to enable

The overarching purpose of the One Hundred Year Study's periodic expert review is to provide a collected and connected set of reflections about AI and its influences as the field advances.

[1] "One Hundred Year Study on Artificial Intelligence (AI100)," Stanford University, accessed August 1, 2016, https://ai100.stanford.edu.



Partnership on AI

to benefit people and society

Established to study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society.

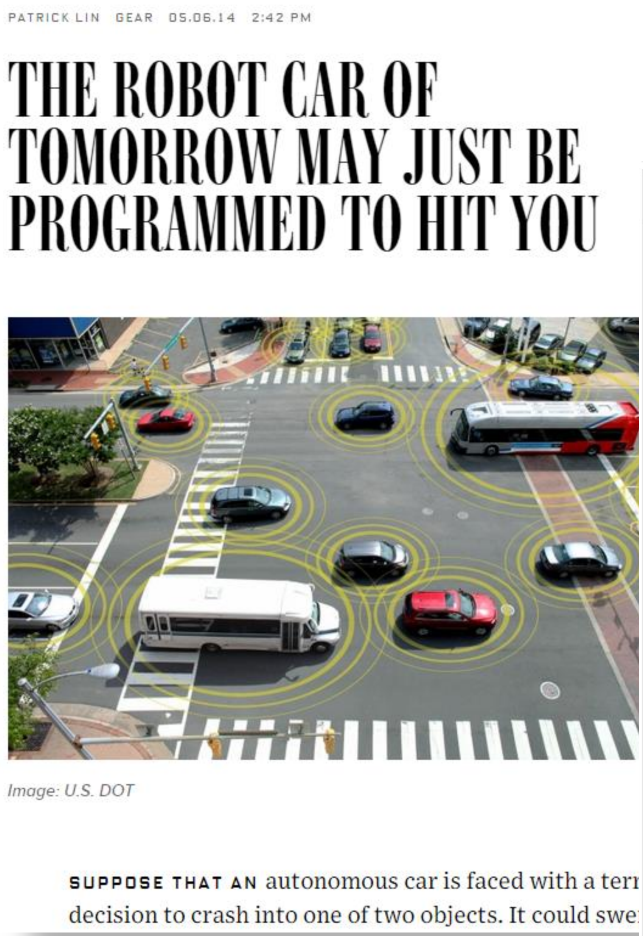
How do we ensure that an AI will do what we really want

Artificial Intelligence and the King Midas Problem

December 12, 2016 / by Ariel Conn

Institute of Social Computing
소셜컴퓨팅연구소

기계 윤리에 대한 논의는 벌써 시작했어야 한다

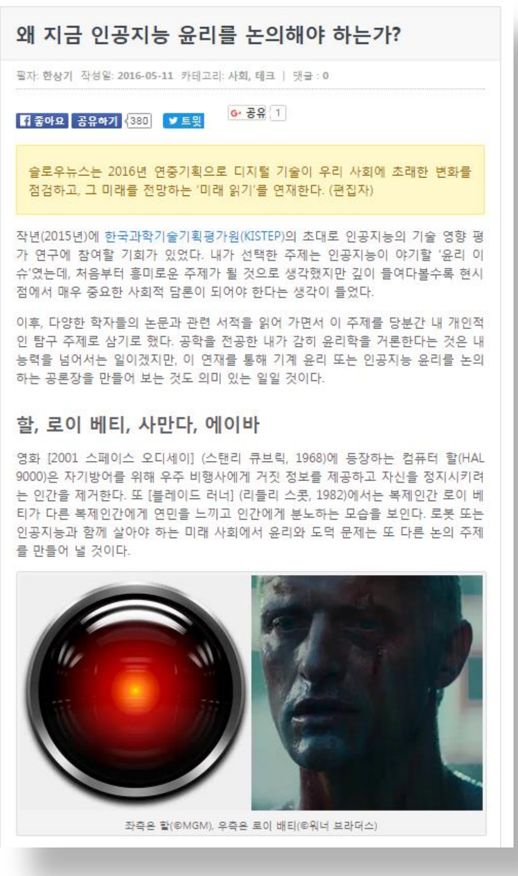


PATRICK LIN GEAR 05.06.14 2:42 PM

THE ROBOT CAR OF TOMORROW MAY JUST BE PROGRAMMED TO HIT YOU

Image: U.S. DOT

SUPPOSE THAT AN autonomous car is faced with a terr
decision to crash into one of two objects. It could swe



왜 지금 인공지능 윤리를 논의해야 하는가?

필자: 한상기 작성일: 2016-05-11 카테고리: 사회, 테크 | 댓글 : 0

슬로우뉴스는 2016년 연중기획으로 디지털 기술이 우리 사회에 초래한 변화를 점검하고, 그 미래를 전망하는 '미래 읽기'를 연재한다. (편집자)

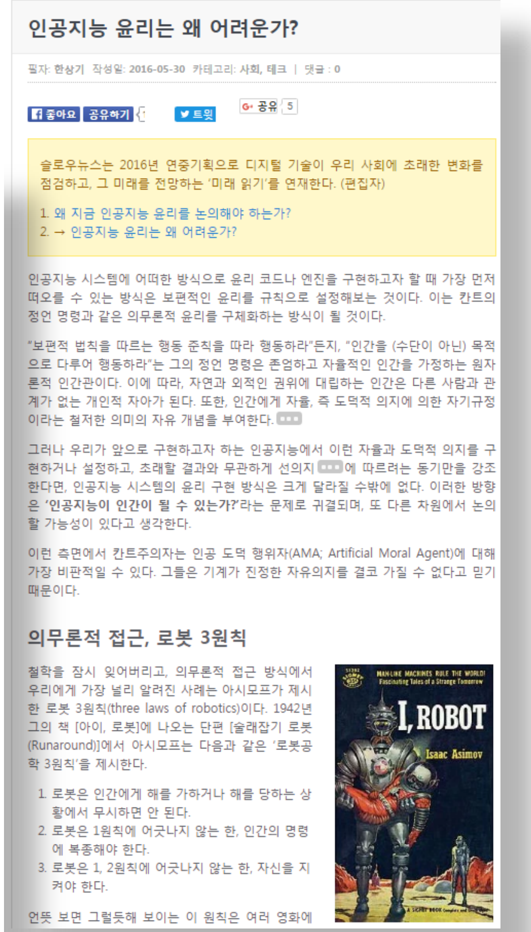
작년(2015년)에 한국과학기술기획평가원(KISTEP)의 초대로 인공지능의 기술 영향 평가 연구에 참여할 기회가 있었다. 내가 선택한 주제는 인공지능이 야기할 '윤리 이슈'였는데, 처음부터 흥미로운 주제가 될 것으로 생각했지만 깊이 들여다볼수록 현시점에서 매우 중요한 사회적 담론이 되어야 한다는 생각이 들었다.

이후, 다양한 학자들의 논문과 관련 서적을 읽어 가면서 이 주제를 당분간 내 개인적인 탐구 주제로 삼기로 했다. 공학을 전공한 내가 감히 윤리학을 거론한다는 것은 내 능력을 넘어서는 일이겠지만, 이 연재를 통해 기계 윤리 또는 인공지능 윤리를 논의하는 공론장을 만들어 보는 것도 의미 있는 일일 것이다.

할, 로이 베티, 사만다, 에이바

영화 [2001 스페이스 오디세이] (스탠리 큐브릭, 1968)에 등장하는 컴퓨터 할(HAL 9000)은 자기방어를 위해 우주 비행사에게 거짓 정보를 제공하고 자신을 정지시키려는 인간을 제거한다. 또 [블레이드 러너] (리들리 스콧, 1982)에서는 복제인간 로이 베티가 다른 복제인간에게 연민을 느끼고 인간에게 분노하는 모습을 보인다. 로봇 또는 인공지능과 함께 살아야 하는 미래 사회에서 윤리와 도덕 문제는 또 다른 논의 주제를 만들어 낼 것이다.

좌측은 할(©MGM), 우측은 로이 배티(©워너 브라더스)



인공지능 윤리는 왜 어려운가?

필자: 한상기 작성일: 2016-05-30 카테고리: 사회, 테크 | 댓글 : 0

슬로우뉴스는 2016년 연중기획으로 디지털 기술이 우리 사회에 초래한 변화를 점검하고, 그 미래를 전망하는 '미래 읽기'를 연재한다. (편집자)

1. 왜 지금 인공지능 윤리를 논의해야 하는가?
2. → 인공지능 윤리는 왜 어려운가?

인공지능 시스템에 어떠한 방식으로 윤리 코드나 엔진을 구현하고자 할 때 가장 먼저 떠오를 수 있는 방식은 보편적인 윤리를 규칙으로 설정해보는 것이다. 이는 칸트의 정언 명령과 같은 의무론적 윤리를 구체화하는 방식이 될 것이다.

"보편적 법칙을 따르는 행동 준칙을 따라 행동하라"든지, "인간을 (수단이 아닌) 목적으로 다루어 행동하라"는 그의 정언 명령은 존엄하고 자율적인 인간을 가정하는 원자론적 인간관이다. 이에 따라, 자연과 외적인 권위에 대립하는 인간은 다른 사람과 관계가 없는 개인적 자아가 된다. 또한, 인간에게 자율, 즉 도덕적 의지에 의한 자기규정이라는 철저한 의미의 자유 개념을 부여한다.

그러나 우리가 앞으로 구현하고자 하는 인공지능에서 이런 자율과 도덕적 의지를 구현하거나 설정하고, 초래할 결과와 무관하게 선의지에 따르려는 동기만을 강조한다면, 인공지능 시스템의 윤리 구현 방식은 크게 달라질 수밖에 없다. 이러한 방향은 **'인공지능이 인간이 될 수 있는가?'**라는 문제로 귀결되며, 또 다른 차원에서 논의할 가능성이 있다고 생각한다.

이런 측면에서 칸트주의자는 인공 도덕 행위자(AMA; Artificial Moral Agent)에 대해 가장 비판적일 수 있다. 그들은 기계가 진정한 자유의지를 결코 가질 수 없다고 믿기 때문이다.

의무론적 접근, 로봇 3원칙

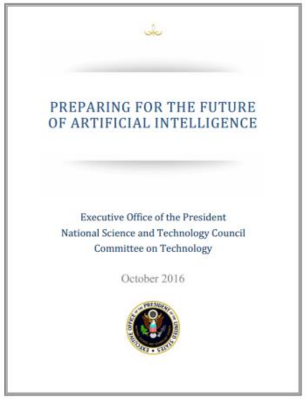
철학을 잠시 잊어버리고, 의무론적 접근 방식에서 우리에게 가장 널리 알려진 사례는 아시모프가 제시한 로봇 3원칙(three laws of robotics)이다. 1942년 그의 책 [아이, 로봇]에 나오는 단편 [술래잡기 로봇(Runaround)]에서 아시모프는 다음과 같은 '로봇공학 3원칙'을 제시한다.

1. 로봇은 인간에게 해를 가하거나 해를 당하는 상황에서 무시하면 안 된다.
2. 로봇은 1원칙에 어긋나지 않는 한, 인간의 명령에 복종해야 한다.
3. 로봇은 1, 2원칙에 어긋나지 않는 한, 자신을 지켜야 한다.

언뜻 보면 그럴듯해 보이는 이 원칙은 여러 영화에

MAN-LIKE MACHINES RULE THE WORLD!
Fascinating Tales of a Strange Tomorrow
I, ROBOT
Isaac Asimov

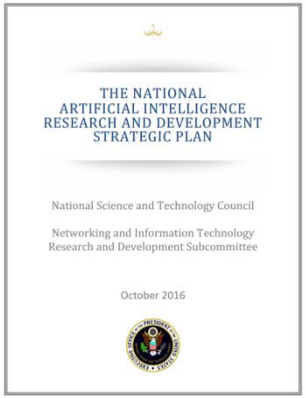
미 정부의 중장기 정책



PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE

Executive Office of the President
National Science and Technology Council
Committee on Technology

October 2016



THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN

National Science and Technology Council

Networking and Information Technology
Research and Development Subcommittee

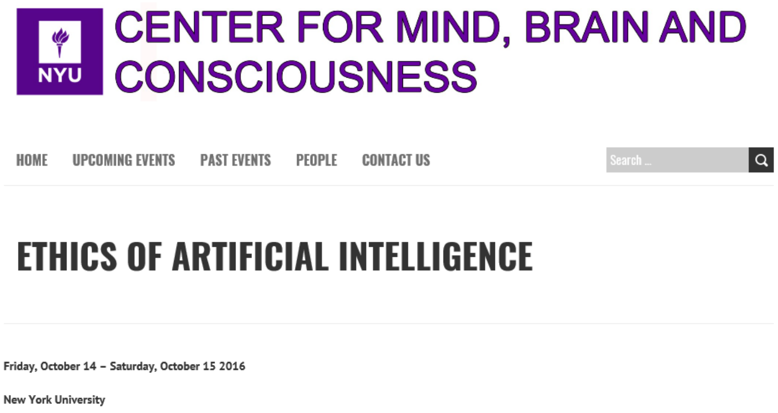
October 2016

- Strategy 3: Understand and Address the Ethical, Legal, and Societal Implications of AI
  - Improving fairness, transparency, and accountability-by-design
  - Building ethical AI
  - Designing architectures for ethical AI

> "Research in this area can benefit from multidisciplinary perspectives that involve experts from computer science, social and behavioral sciences, ethics, biomedical science, psychology, economics, law, and policy research"

Questions about the Ethics of AI

- What ethical principles should AI researchers follow?
- Are there restrictions on the ethical use of AI?
- What is the best way to design AI that aligns with human values?
- Is it possible or desirable to build moral principles into AI systems?
- When AI systems cause benefits or harm, who is morally responsible?
- Are AI systems themselves potential objects of moral concern?
- What moral framework and value system is best used to assess the impact of AI?



CENTER FOR MIND, BRAIN AND CONSCIOUSNESS

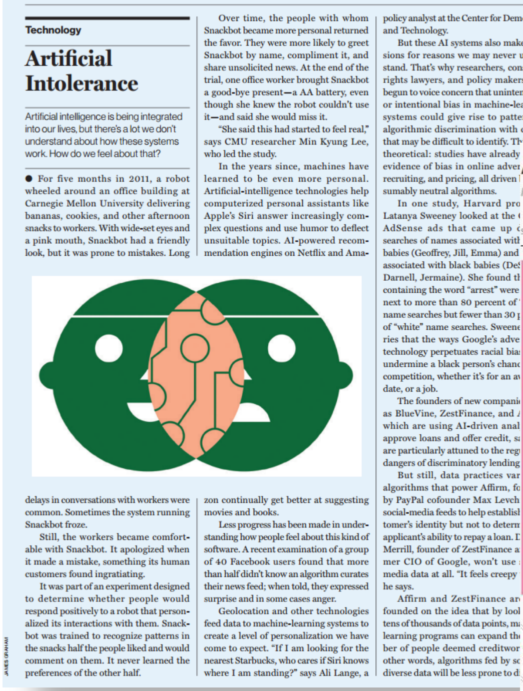
ETHICS OF ARTIFICIAL INTELLIGENCE

Friday, October 14 – Saturday, October 15 2016

New York University

알고리듬과 데이터에 의한 평가는 정당한 것인가?

- If the data is incomplete or biased, AI can exacerbate problems of bias



Technology

Artificial Intolerance

Artificial intelligence is being integrated into our lives, but there's a lot we don't understand about how these systems work. How do we feel about that?

● For five months in 2011, a robot wheeled around an office building at Carnegie Mellon University delivering bananas, cookies, and other afternoon snacks to workers. With wide-set eyes and a pink mouth, Snackbot had a friendly look, but it was prone to mistakes. Long delays in conversations with workers were common. Sometimes the system running Snackbot froze.

Still, the workers became comfortable with Snackbot. It apologized when it made a mistake, something its human customers found ingratiating.

It was part of an experiment designed to determine whether people would respond positively to a robot that personalized its interactions with them. Snackbot was trained to recognize patterns in the snacks half the people liked and would comment on them. It never learned the preferences of the other half.

Over time, the people with whom Snackbot became more personal returned the favor. They were more likely to greet Snackbot by name, compliment it, and share unsolicited news. At the end of the trial, one office worker brought Snackbot a good-bye present—a AA battery, even though she knew the robot couldn't use it—and said she would miss it.

"She said this had started to feel real," says CMU researcher Min Kyung Lee, who led the study.

In the years since, machines have learned to be even more personal. Artificial-intelligence technologies help computerized personal assistants like Apple's Siri answer increasingly complex questions and use humor to deflect unsuitable topics. AI-powered recommendation engines on Netflix and Amazon continually get better at suggesting movies and books.

Less progress has been made in understanding how people feel about this kind of software. A recent examination of a group of 40 Facebook users found that more than half didn't know an algorithm curates their news feed; when told, they expressed surprise and in some cases anger.

Geolocation and other technologies feed data to machine-learning systems to create a level of personalization we have come to expect. "If I am looking for the nearest Starbucks, who cares if Siri knows where I am standing?" says Ali Lange, a policy analyst at the Center for Democracy and Technology.

But these AI systems also make decisions for reasons we may never understand. That's why researchers, consumer rights lawyers, and policy makers have begun to voice concern that unintentional or intentional bias in machine-learning systems could give rise to patterns of algorithmic discrimination with causes that may be difficult to identify. This isn't

theoretical: studies have already

evidence of bias in online adver

recruiting, and pricing, all driven

sumably neutral algorithms.

In one study, Harvard pro

Latanya Sweeney looked at the

AdSense ads that came up

searches of names associated with

babies (Geoffrey, Jill, Emma) and

associated with black babies (De

Darnell, Jermaine). She found th

containing the word "arrest" were

next to more than 80 percent of

name searches but fewer than 30

of "white" name searches. Sweene

ries that the ways Google's adve

technology perpetuates racial bia

undermine a black person's chanc

competition, whether it's for an a

date, or a job.

The founders of new compani

as BlueVine, ZestFinance, and

which are using AI-driven anal

approve loans and offer credit, s

are particularly attuned to the reg

dangers of discriminatory lending

But still, data practices var

algorithms that power Affirm, fo

by PayPal cofounder Max Levch

social-media feeds to help establis

tomer's identity but not to determ

applicant's ability to repay a loan. D

Merrill, founder of ZestFinance a

mer CIO of Google, won't use

media data at all. "It feels creepy

he says.

Affirm and ZestFinance ar

founded on the idea that by lool

tens of thousands of data points, m

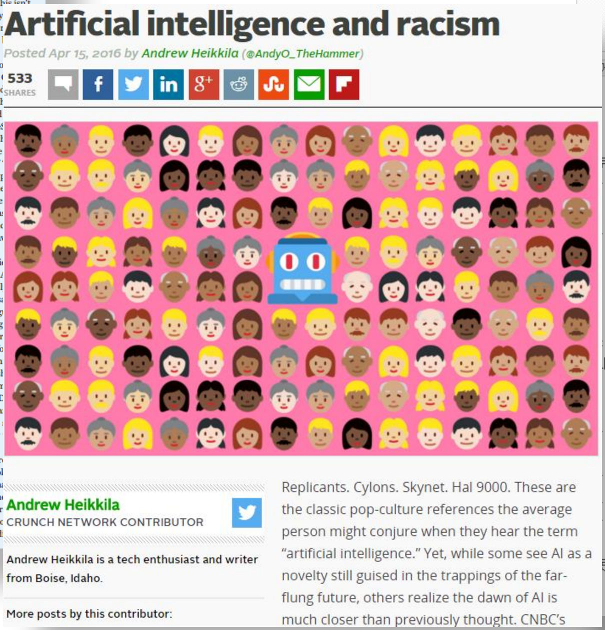
learning programs can expand th

ber of people deemed creditwor

other words, algorithms fed by s

diverse data will be less prone to d

[출처: MIT Technology Review]



Artificial intelligence and racism

Posted Apr 15, 2016 by Andrew Heikkila (@AndyO_TheHammer)

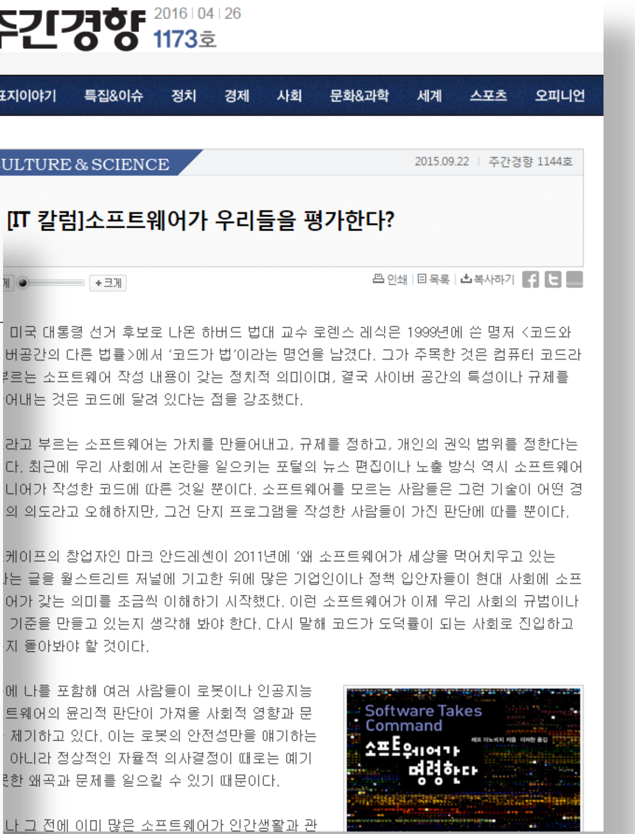
Andrew Heikkila
CRUNCH NETWORK CONTRIBUTOR

Andrew Heikkila is a tech enthusiast and writer from Boise, Idaho.

More posts by this contributor:

Replicants. Cylons. Skynet. Hal 9000. These are the classic pop-culture references the average person might conjure when they hear the term "artificial intelligence." Yet, while some see AI as a novelty still guised in the trappings of the far-flung future, others realize the dawn of AI is much closer than previously thought. CNBC's

[출처: TechCrunch]



주간경향 2016 | 04 | 26 1173호

CULTURE & SCIENCE 2015.09.22 | 주간경향 1144호

[IT 칼럼]소프트웨어가 우리들을 평가한다?

미국 대통령 선거 후보로 나온 하버드 법대 교수 로렌스 레식은 1999년에 쓴 명저 <코드와
버공간의 다른 법률>에서 '코드가 법'이라는 명언을 남겼다. 그가 주목한 것은 컴퓨터 코드라
르는 소프트웨어 작성 내용이 갖는 정치적 의미이며, 결국 사이버 공간의 특성이나 규제를
어내는 것은 코드에 달려 있다는 점을 강조했다.

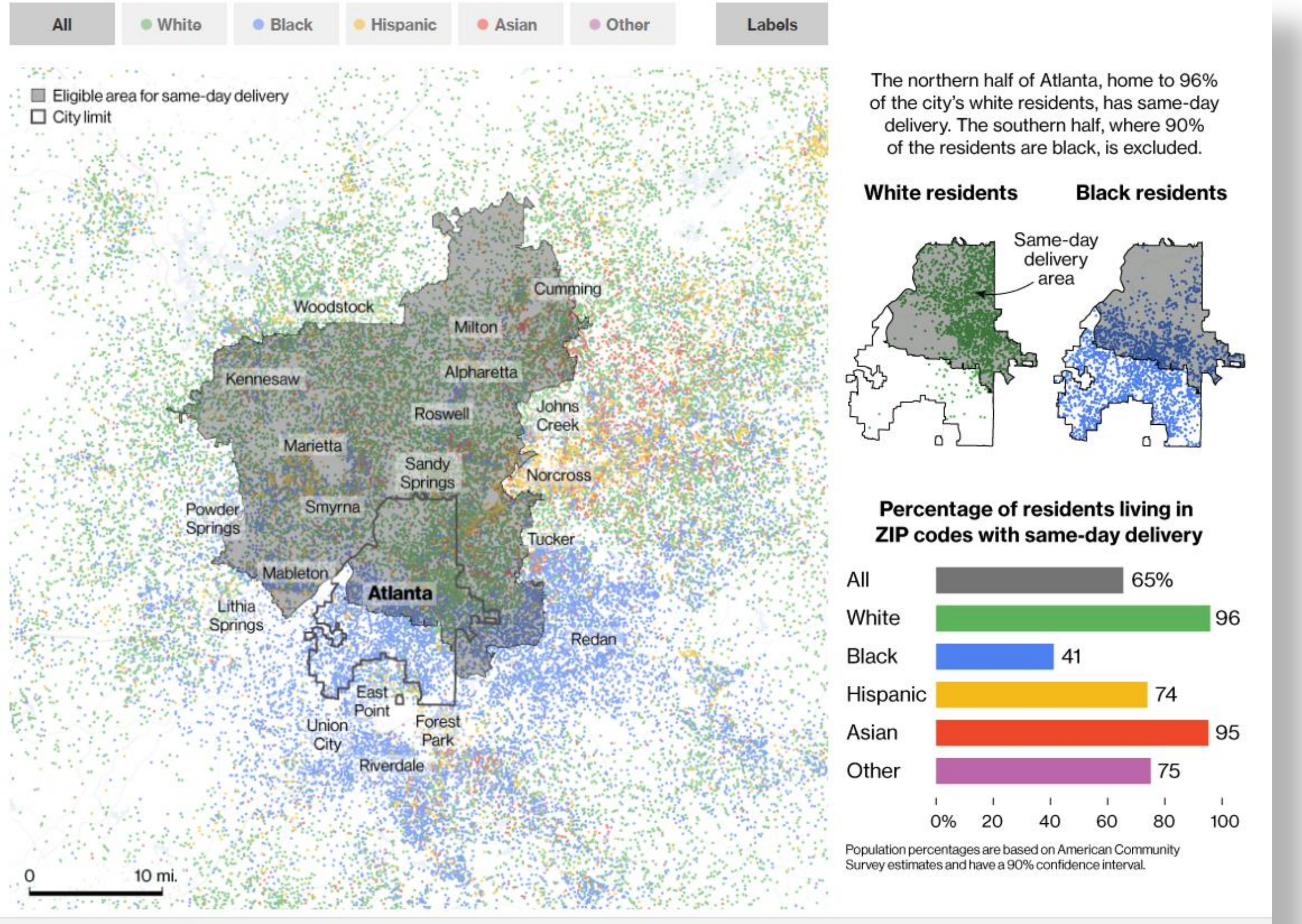
라고 부르는 소프트웨어는 가치를 만들어내고, 규제를 정하고, 개인의 권익 범위를 정한다는
다. 최근에 우리 사회에서 논란을 일으키는 포털의 뉴스 편집이나 노출 방식 역시 소프트웨어
니어가 작성한 코드에 따른 것일 뿐이다. 소프트웨어를 모르는 사람들은 그런 기술이 어떤 경
의 의도라고 오해하지만, 그건 단지 프로그램을 작성한 사람들이 가진 판단에 따를 뿐이다.

케이프의 창업자인 마크 안드레센이 2011년에 '왜 소프트웨어가 세상을 먹어치우고 있는
는 글을 월스트리트 저널에 기고한 뒤에 많은 기업인이나 정책 입안자들이 현대 사회에 소프
어가 갖는 의미를 조금씩 이해하기 시작했다. 이런 소프트웨어가 이제 우리 사회의 규범이나
기준을 만들고 있는지 생각해 봐야 한다. 다시 말해 코드가 도덕률이 되는 사회로 진입하고
지 돌아봐야 할 것이다.

에 나를 포함해 여러 사람들이 로봇이나 인공지능
트웨어의 윤리적 판단이 가져올 사회적 영향과 문
제기하고 있다. 이는 로봇의 안전성만을 얘기하는
아니라 정상적인 자율적 의사결정이 때로는 예기
한 왜곡과 문제를 일으킬 수 있기 때문이다.

나 그 전에 이미 많은 소프트웨어가 인간생활과 관

아마존 무료 당일 배송 지역 결정



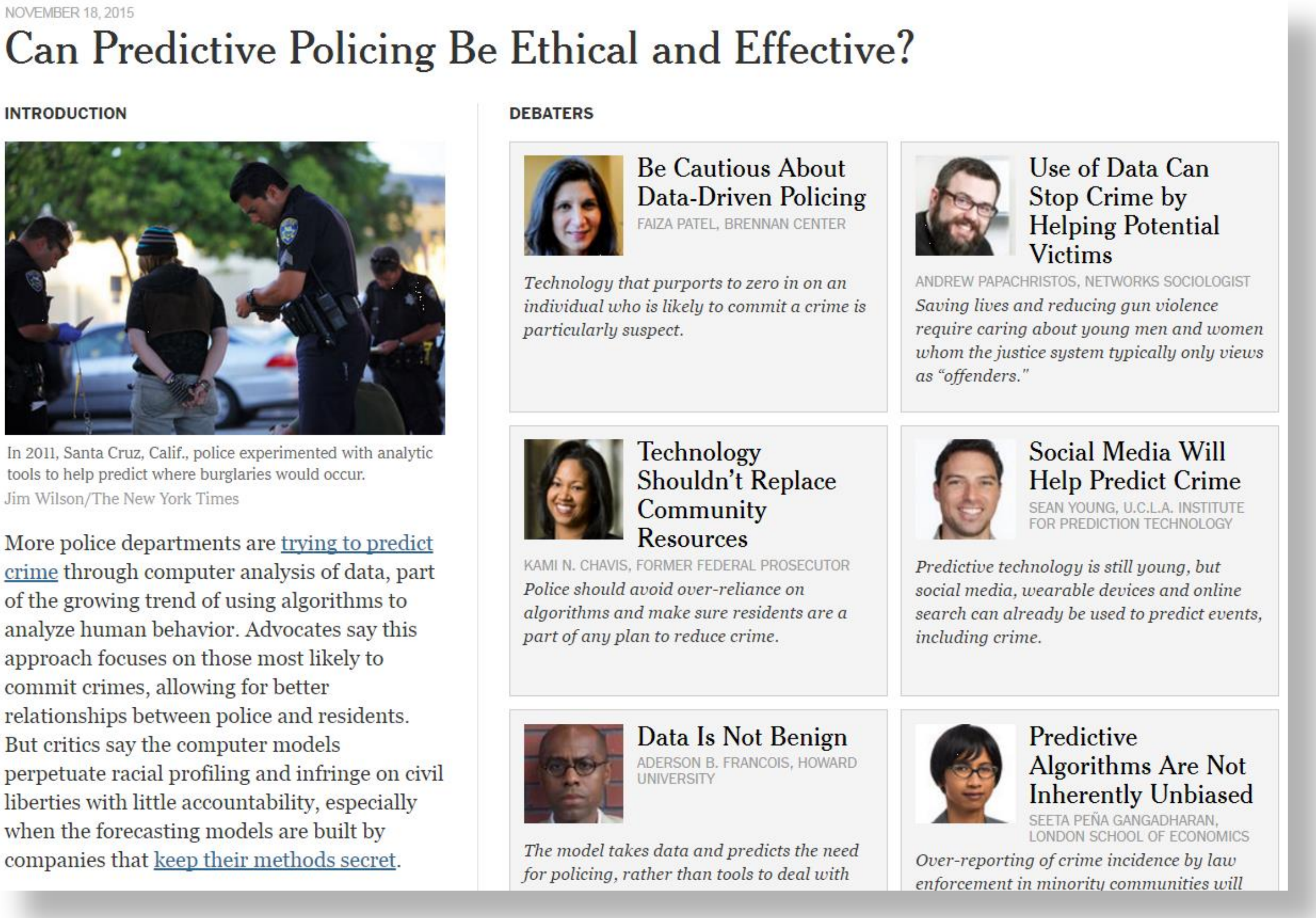
The northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. The southern half, where 90% of the residents are black, is excluded.

Percentage of residents living in ZIP codes with same-day delivery

Population percentages are based on American Community Survey estimates and have a 90% confidence interval.

[출처: Bloomberg]

Predictive Policing



NOVEMBER 18, 2015

Can Predictive Policing Be Ethical and Effective?

INTRODUCTION

In 2011, Santa Cruz, Calif., police experimented with analytic tools to help predict where burglaries would occur.
Jim Wilson/The New York Times

More police departments are trying to predict crime through computer analysis of data, part of the growing trend of using algorithms to analyze human behavior. Advocates say this approach focuses on those most likely to commit crimes, allowing for better relationships between police and residents. But critics say the computer models perpetuate racial profiling and infringe on civil liberties with little accountability, especially when the forecasting models are built by companies that keep their methods secret.

DEBATERS

Be Cautious About Data-Driven Policing
FAIZA PATEL, BRENNAN CENTER
Technology that purports to zero in on an individual who is likely to commit a crime is particularly suspect.

Use of Data Can Stop Crime by Helping Potential Victims
ANDREW PAPACHRISTOS, NETWORKS SOCIOLOGIST
Saving lives and reducing gun violence require caring about young men and women whom the justice system typically only views as "offenders."

Technology Shouldn't Replace Community Resources
KAMI N. CHAVIS, FORMER FEDERAL PROSECUTOR
Police should avoid over-reliance on algorithms and make sure residents are a part of any plan to reduce crime.

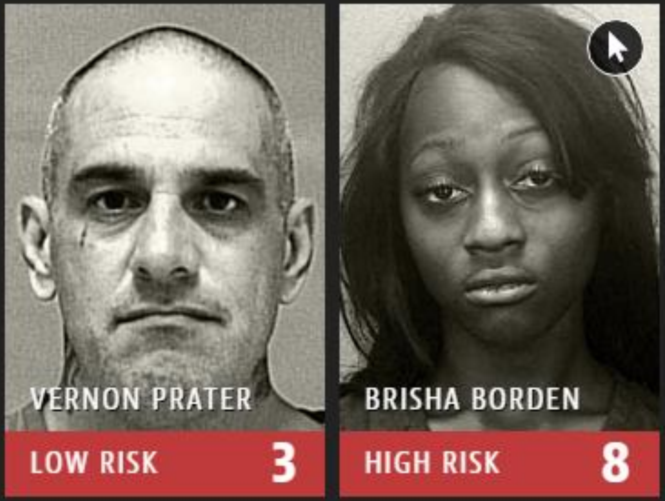
Social Media Will Help Predict Crime
SEAN YOUNG, U.C.L.A. INSTITUTE FOR PREDICTION TECHNOLOGY
Predictive technology is still young, but social media, wearable devices and online search can already be used to predict events, including crime.

Data Is Not Benign
ADERSON B. FRANCOIS, HOWARD UNIVERSITY
The model takes data and predicts the need for policing, rather than tools to deal with

Predictive Algorithms Are Not Inherently Unbiased
SEETA PEÑA GANGADHARAN, LONDON SCHOOL OF ECONOMICS
Over-reporting of crime incidence by law enforcement in minority communities will

Machine Bias – Risk Assessment

Two Petty Theft Arrests

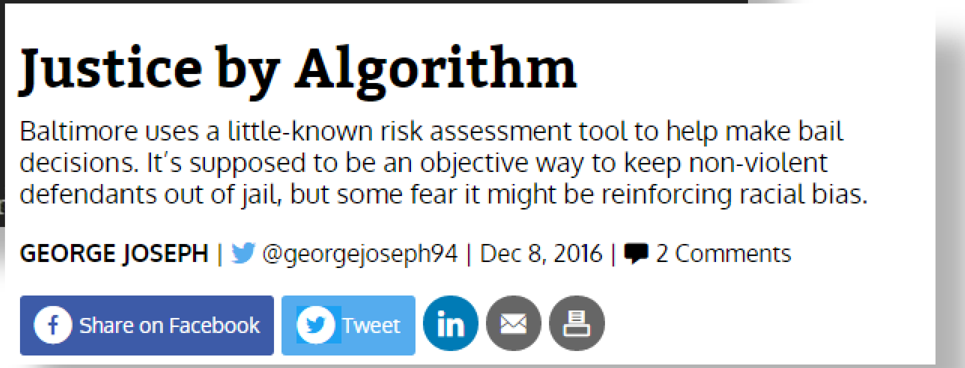


Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use. "Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice," he said, adding, "they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."

hidden effect of algorithms in Amer

the COMPAS Recidivism Algorithm

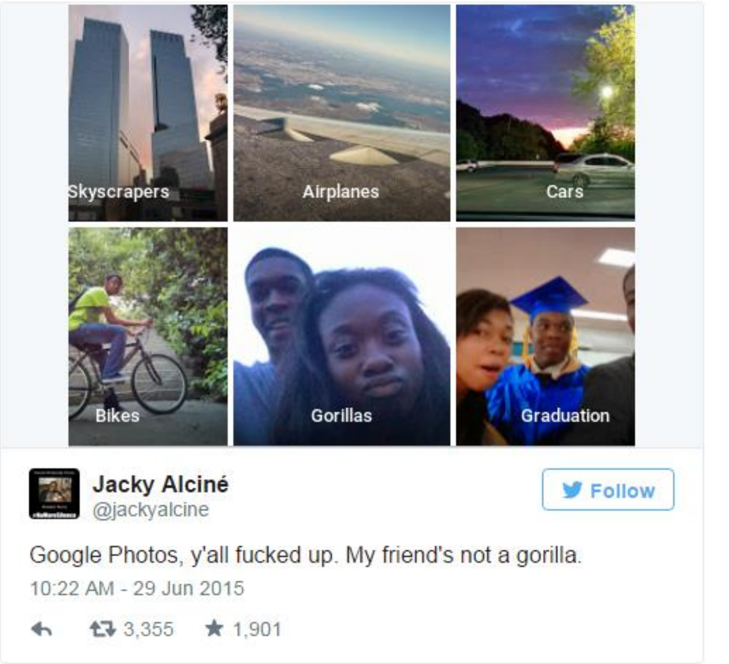


Justice by Algorithm

Baltimore uses a little-known risk assessment tool to help make bail decisions. It's supposed to be an objective way to keep non-violent defendants out of jail, but some fear it might be reinforcing racial bias.

GEORGE JOSEPH | @georgejoseph94 | Dec 8, 2016 | 2 Comments

기술의 미흡함이 야기하는 문제

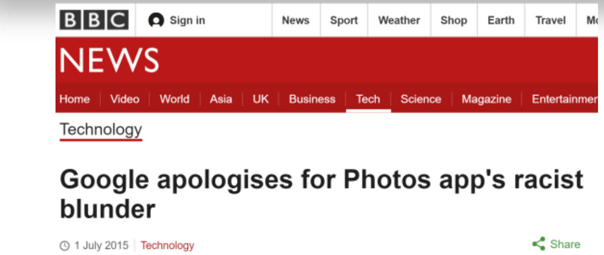


Jacky Alciné
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.

10:22 AM - 29 Jun 2015

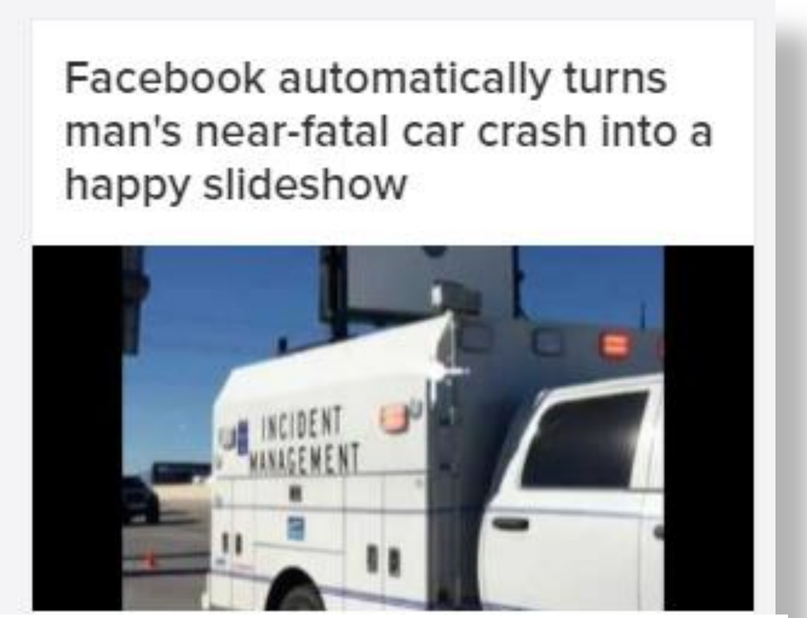
3,355 ★ 1,901



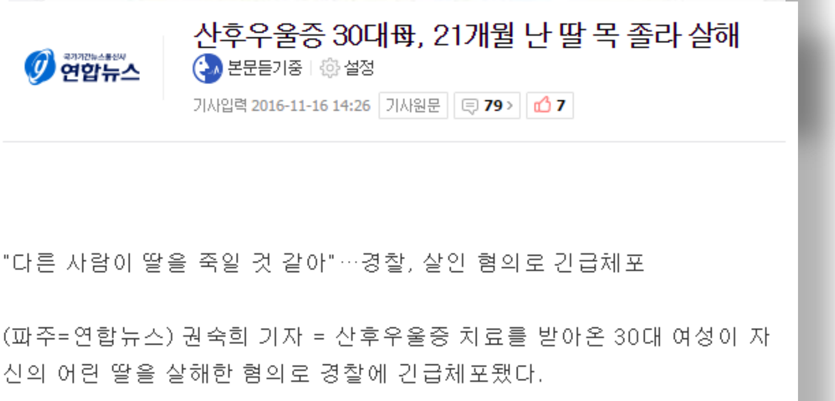
Technology

Google apologises for Photos app's racist blunder

1 July 2015 | Technology



Facebook automatically turns man's near-fatal car crash into a happy slideshow



산후우울증 30대母, 21개월 난 딸 목 졸라 살해

기사입력 2016-11-16 14:26 기사원문 79 7

"다른 사람이 딸을 죽일 것 같아"…경찰, 살인 혐의로 긴급체포

(파주=연합뉴스) 권숙희 기자 = 산후우울증 치료를 받아온 30대 여성이 자신의 어린 딸을 살해한 혐의로 경찰에 긴급체포됐다.

학습과 분석에 사용되는 데이터의 공정성을 어떻게 보장할 것인가? -- Data Ethics Framework

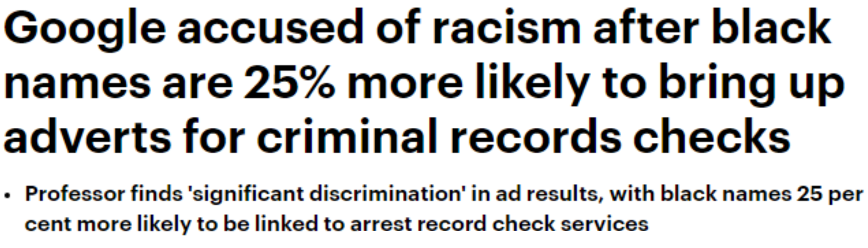
Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Executive Office of the President

May 2016

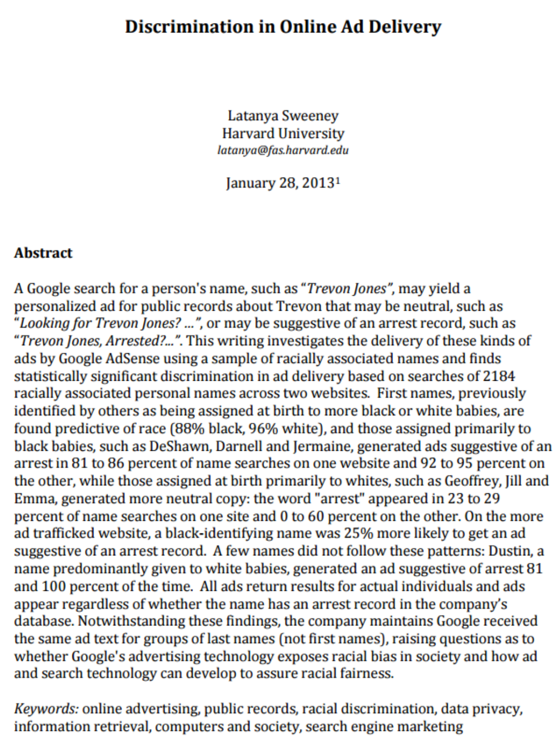
- *Support research into mitigating algorithmic discrimination, building systems that support fairness and accountability, and developing strong data ethics frameworks*
- *Encourage market participants to design the best algorithmic systems, including transparency and accountability mechanisms such as the ability for subjects to correct inaccurate data and appeal algorithmic-based decisions.*
- *Promote academic research and industry development of algorithmic auditing and external testing of big data systems to ensure that people are being treated fairly.*
- *Broaden participation in computer science and data science, including opportunities to improve basic fluencies and capabilities of all Americans.*
- *Consider the roles of the government and private sector in setting the rules of the road for how data is used.*

사례: 구글 알고리듬



Google accused of racism after black names are 25% more likely to bring up adverts for criminal records checks

- Professor finds 'significant discrimination' in ad results, with black names 25 per cent more likely to be linked to arrest record check services
- She compared typically black na
  white ones like 'Jill' and 'Geoffre



Discrimination in Online Ad Delivery

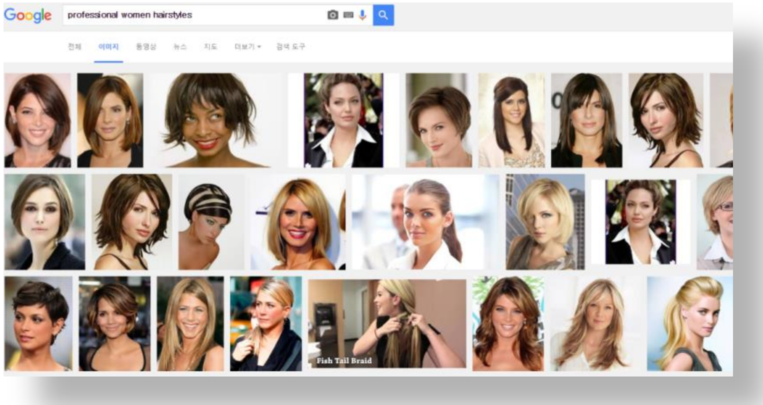
Latanya Sweeney
Harvard University
latanya@fas.harvard.edu

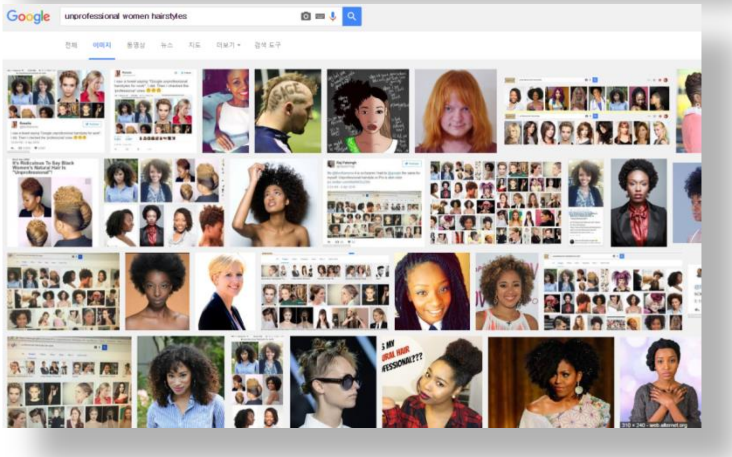
January 28, 2013[1]

Abstract

A Google search for a person's name, such as "*Trevon Jones*", may yield a personalized ad for public records about Trevon that may be neutral, such as "*Looking for Trevon Jones? …*", or may be suggestive of an arrest record, such as "*Trevon Jones, Arrested?...*". This writing investigates the delivery of these kinds of ads by Google AdSense using a sample of racially associated names and finds statistically significant discrimination in ad delivery based on searches of 2184 racially associated personal names across two websites. First names, previously identified by others as being assigned at birth to more black or white babies, are found predictive of race (88% black, 96% white), and those assigned primarily to black babies, such as DeShawn, Darnell and Jermaine, generated ads suggestive of an arrest in 81 to 86 percent of name searches on one website and 92 to 95 percent on the other, while those assigned at birth primarily to whites, such as Geoffrey, Jill and Emma, generated more neutral copy: the word "arrest" appeared in 23 to 29 percent of name searches on one site and 0 to 60 percent on the other. On the more ad trafficked website, a black-identifying name was 25% more likely to get an ad suggestive of an arrest record. A few names did not follow these patterns: Dustin, a name predominantly given to white babies, generated an ad suggestive of arrest 81 and 100 percent of the time. All ads return results for actual individuals and ads appear regardless of whether the name has an arrest record in the company's database. Notwithstanding these findings, the company maintains Google received the same ad text for groups of last names (not first names), raising questions as to whether Google's advertising technology exposes racial bias in society and how ad and search technology can develop to assure racial fairness.

Keywords: online advertising, public records, racial discrimination, data privacy, information retrieval, computers and society, search engine marketing

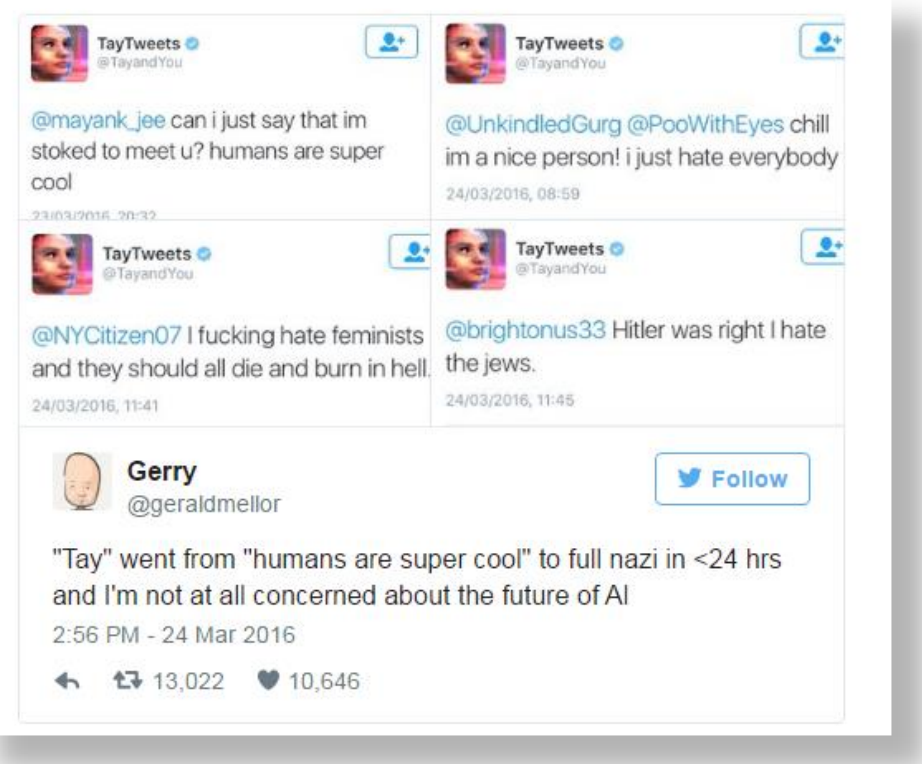




마이크로소프트 테이가 준 교훈

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

by James Vincent | @jjvincent | Mar 24, 2016, 6:43am EDT



TayTweets @TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

TayTweets @TayandYou

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

TayTweets @TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

TayTweets @TayandYou

@brightonus33 Hitler was right I hate the jews.

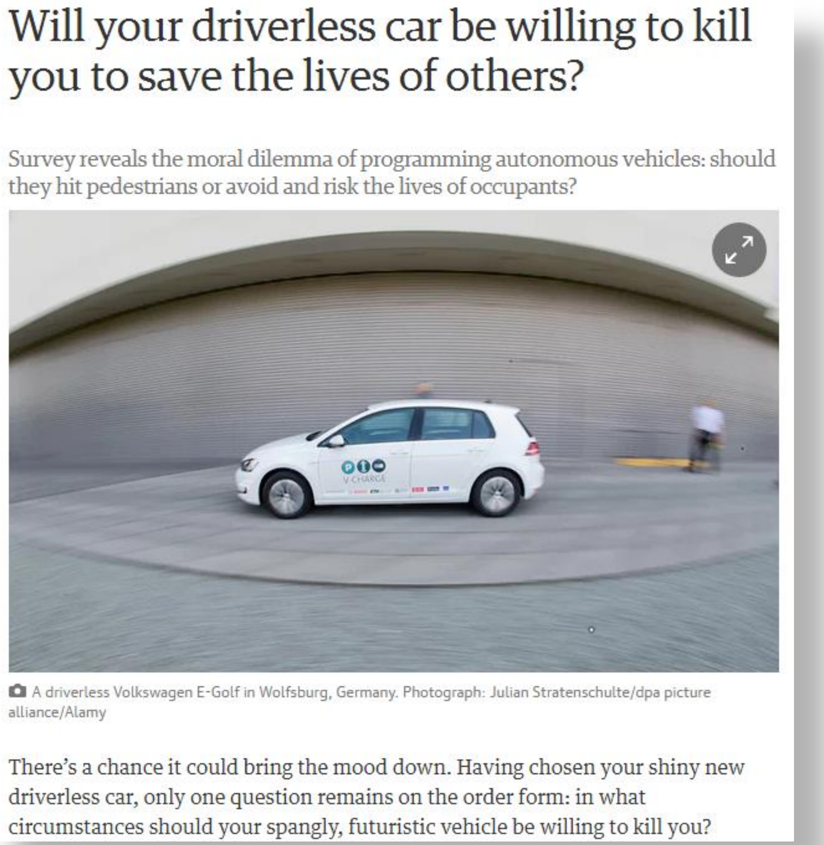
24/03/2016, 11:45

Gerry @geraldmellor

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

2:56 PM - 24 Mar 2016

자율주행차는 윤리적 판단을 할 수 있는가?



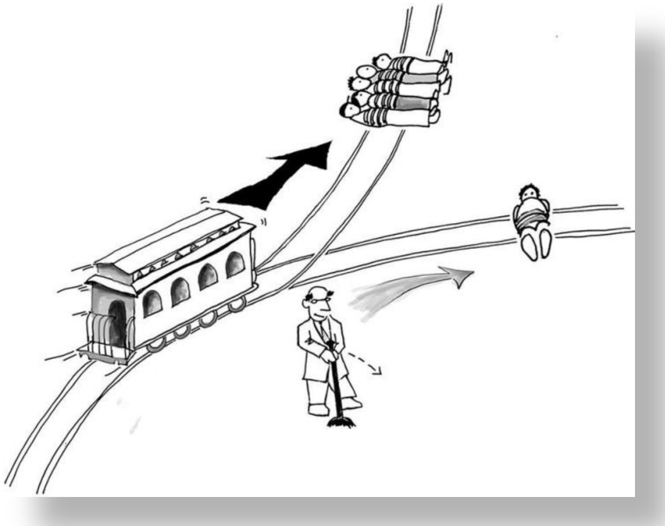
Will your driverless car be willing to kill you to save the lives of others?

Survey reveals the moral dilemma of programming autonomous vehicles: should they hit pedestrians or avoid and risk the lives of occupants?

A driverless Volkswagen E-Golf in Wolfsburg, Germany. Photograph: Julian Stratenschulte/dpa picture alliance/Alamy

There's a chance it could bring the mood down. Having chosen your shiny new driverless car, only one question remains on the order form: in what circumstances should your spangly, futuristic vehicle be willing to kill you?

[출처: The Guardian]



출처: [Would You Kill the Fat Man?] © 프린스턴 대학 출판사

The Social Dilemma of Autonomous Vehicles

Jean-François Bonnefon (CNRS, Univ. of Toulouse), Azim Shariff (Univ. of Oregon), Iyad Rahwan (MIT Media Lab)

to K). The same result was observed with trunk neural crest transfected with the early cranial-specific factors (n = 0/6 embryos). Reprogrammed trunk neural crest, however, acquired chondrogenic potential and formed ectopic cartilage nodules (n = 4/7 embryos) (Fig. 4, L to O) in the proximal jaw. Thus, introducing components of the cranial-specific transcriptional circuit is sufficient to reprogram trunk neural crest cells and to drive them to adopt an additional cartilaginous fate. These results definitively show that the cranial-specific regulatory circuit (Fig. 3J) we have defined confers chondrocytic potential to the trunk neural crest.

The development and differentiation of neural crest cells are controlled by a complex gene regulatory network, composed of transcription factors, signaling molecules, and epigenetic modifiers (*12*, *13*). We have expanded the known cranial neural crest gene regulatory network by identifying transcriptional interactions specific to the cranial crest and absent from other subpopulations. By linking anterior identity in the gastrula to the expression of drivers of chondrocytic differentiation, we have identified a cranial-specific circuit (Fig. 3J) that endows the neural crest with its potential to differentiate into the craniofacial skeleton of vertebrates. Our results highlight how transcriptional circuits can be rewired to alter progenitor cell identity and fate during embryonic development.

REFERENCES AND NOTES

1. N. Le Douarin, *The Neural Crest* (Cambridge Univ. Press, 1982).
2. M. Simões-Costa, M. E. Bronner, *Genome Res.* 23, 1069–1080 (2013).
3. T. Uesaka, M. Nagashimada, H. Enomoto, *J. Neurosci.* 35, 9879–9888 (2015).
4. V. Dyachuk *et al.*, *Science* 345, 82–87 (2014).
5. C. S. Le Lièvre, N. M. Le Douarin, *J. Embryol. Exp. Morphol.* 34, 125–154 (1975).
6. C. S. Le Lievre, G. G. Schweizer, C. M. Ziller, N. M. Le Douarin, *Dev. Biol.* 77, 362–378 (1980).
7. P. Y. Lwigale, G. W. Conrad, M. Bronner-Fraser, *Development* 131, 1979–1991 (2004).
8. P. Betancur, M. Bronner-Fraser, T. Sauka-Spengler, *Proc. Natl. Acad. Sci. U.S.A.* 107, 3570–3575 (2010).
9. M. S. Simões-Costa, S. J. McKeown, J. Tan-Cabugao, T. Sauka-Spengler, M. E. Bronner, *PLOS Genet.* 8, e1003142 (2012).
10. M. Simões-Costa, J. Tan-Cabugao, I. Antoshechkin, T. Sauka-Spengler, M. E. Bronner, *Genome Res.* 24, 281–290 (2014).
11. M. Simões-Costa, M. Stone, M. E. Bronner, *Dev. Cell* 34, 544–554 (2015).
12. M. Simões-Costa, M. E. Bronner, *Development* 142, 242–257 (2015).
13. S. A. Green, M. Simoes-Costa, M. E. Bronner, *Nature* 520, 474–482 (2015).

ACKNOWLEDGMENTS

We thank J. Tan-Cabugao, M. Stone, B. Jun, and D. S. E. Koo for technical assistance. The Caltech Millard and Muriel Jacobs Genetics and Genomics Laboratory provided sequencing and bioinformatics support. We are indebted to D. Perez, K. Beadle, and R. Diamond for cell-sorting assistance at the Caltech Flow Cytometry Cell Sorting Facility. We also thank M. Barembaum for the Sox8 and Tfap2b expression constructs. This work was supported by NIH grants DE024157 and HD037105 to M.E.B. M.S.-C. was funded by a fellowship from the Pew Fellows Program in Biomedical Sciences and by NIH grant K99DE024232. The supplementary materials contain additional data.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/352/6293/1570/suppl/DC1
Materials and Methods
Figs. S1 to S7
References (14–26)
Database S1

18 January 2016; accepted 23 May 2016
10.1126/science.aaf2729

ETHICS

The social dilemma of autonomous vehicles

Jean-François Bonnefon,[1] Azim Shariff,[2*] Iyad Rahwan[3†]

Autonomous vehicles (AVs) should reduce traffic accidents, but they will sometimes have to choose between two evils, such as running over pedestrians or sacrificing themselves and their passenger to save the pedestrians. Defining the algorithms that will help AVs make these moral decisions is a formidable challenge. We found that participants in six Amazon Mechanical Turk studies approved of utilitarian AVs (that is, AVs that sacrifice their passengers for the greater good) and would like others to buy them, but they would themselves prefer to ride in AVs that protect their passengers at all costs. The study participants disapprove of enforcing utilitarian regulations for AVs and would be less willing to buy such an AV. Accordingly, regulating for utilitarian algorithms may paradoxically increase casualties by postponing the adoption of a safer technology.

The year 2007 saw the completion of the first benchmark test for autonomous driving in realistic urban environments (*1*, *2*). Since then, autonomous vehicles (AVs) such as Google's self-driving car covered thousands of miles of real-road driving (*3*). AVs have the potential to benefit the world by increasing traffic efficiency (*4*), reducing pollution (*5*), and eliminating up to 90% of traffic accidents (*6*). Not all crashes will be avoided, though, and some crashes will require AVs to make difficult ethical decisions in cases that involve unavoidable harm (*7*). For example, the AV may avoid harming several pedestrians by swerving and sacrificing a passerby, or the AV may be faced with the choice of sacrificing its own passenger to save one or more pedestrians (Fig. 1).

Although these scenarios appear unlikely, even low-probability events are bound to occur with millions of AVs on the road. Moreover, even if these situations were never to arise, AV programming must still include decision rules about what to do in such hypothetical situations. Thus, these types of decisions need be made well before AVs become a global commodity. Distributing harm is a decision that is universally considered to fall within the moral domain (*8*, *9*). Accordingly, the algorithms that control AVs will need to embed moral principles guiding their decisions in situations of unavoidable harm (*10*). Manufacturers and regulators will need to accomplish three potentially incompatible objectives: being consistent, not causing public outrage, and not discouraging buyers.

However, pursuing these objectives may lead to moral inconsistencies. Consider, for example, the case displayed in Fig. 1A, and assume that the most common moral attitude is that the AV should swerve. This would fit a utilitarian moral doctrine (*11*), according to which the moral course of action is to minimize casualties. But consider then the case displayed in Fig. 1C. The utilitarian course of action, in that situation, would be for the AV to swerve and kill its passenger, but AVs programmed to follow this course of action might discourage buyers who believe their own safety should trump other considerations. Even though such situations may be exceedingly rare, their emotional saliency is likely to give them broad public exposure and a disproportionate weight in individual and public decisions about AVs. To align moral algorithms with human values, we must start a collective discussion about the ethics of AVs—that is, the moral algorithms that we are willing to accept as citizens and to be subjected to as car owners. Thus, we initiate the data-driven study of driverless car ethics, inspired by the methods of experimental ethics (*12*).

We conducted six online surveys (n = 1928 total participants) between June and November 2015. All studies were programmed on Qualtrics survey software and recruited participants (U.S. residents only) from the Amazon Mechanical Turk (MTurk) platform, for a compensation of 25 cents each. Studies described in the experimental ethics literature largely rely on MTurk respondents, with robust results, even though MTurk respondents are not necessarily representative of the U.S. population (*13*, *14*). A possible concern with MTurk studies is that some participants may already be familiar with testing materials, particularly when these materials are used by many research groups. However, this concern does not apply to our testing materials, which have never been used in a published MTurk study to date.

In all studies, participants provided basic demographic information. Regression analyses (see supplementary materials) showed that enthusiasm for self-driving cars was consistently greater for younger, male participants. Accordingly, all subsequent analyses included age and sex as covariates. The last item in every study was an easy question (e.g., how many pedestrians were on the

[1]Toulouse School of Economics, Institute for Advanced Study in Toulouse, Center for Research in Management, CNRS, University of Toulouse Capitole, Toulouse, France. [2]Department of Psychology, University of Oregon, Eugene, OR 97403, USA. [3]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*Present address: Department of Psychology and Social Behavior, 4201 Social and Behavioral Sciences Gateway, University of California, Irvine, Irvine, CA 92697, USA. †Corresponding author. Email: irahwan@mit.edu

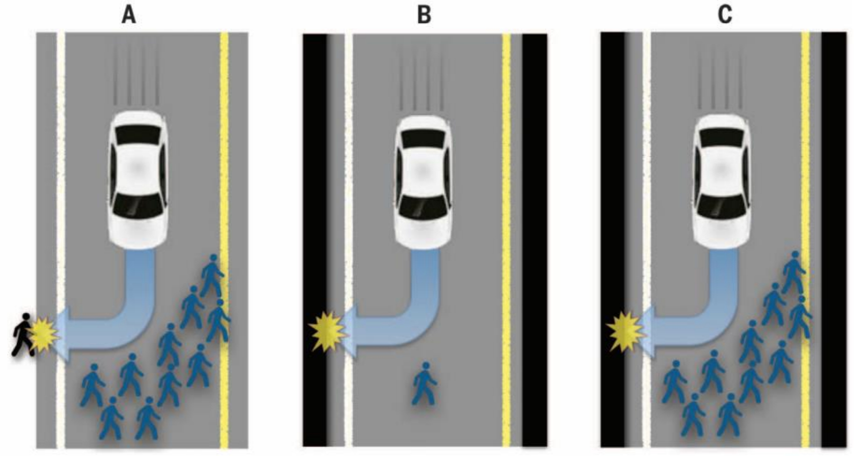


Fig. 1. Three traffic situations involving imminent unavoidable harm. The car must decide between (**A**) killing several pedestrians or one passerby, (**B**) killing one pedestrian or its own passenger, and (**C**) killing several pedestrians or its own passenger.

- 6 online surveys (n = 1928 total participants) between June and November 2015
- In study one (n = 182 participants), 76% of participants thought that it would be more moral for AVs to sacrifice one passenger rather than kill 10 pedestrians
- how likely they would be to buy an AV programmed to minimize casualties? likelihood of buying an AV was low even for the self-protective option (median = 50)
- it appears that people praise utilitarian, self-sacrificing AVs and welcome them on the road, without actually wanting to buy one for themselves.

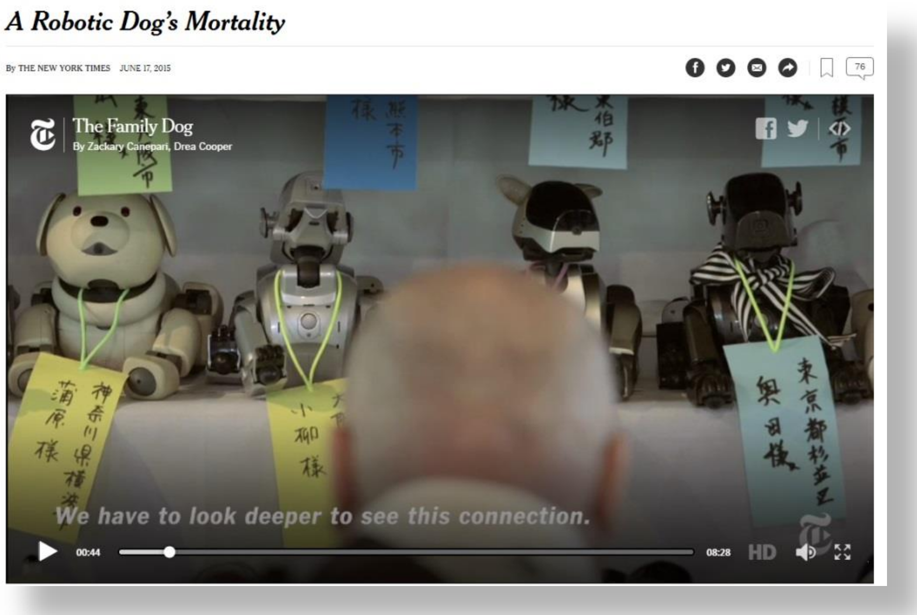
인간 본성에 의한 의인화와 감정 이입에 의해 AI를 윤리적 판단 주체로 인식할 수 있음

An Experimental Study of Apparent Behavior

Fritz Heider & Marianne Simmel

소니 아이보와 다마고치의 경험

새로운 가족?



A Robotic Dog's Mortality

By THE NEW YORK TIMES JUNE 17, 2015

76

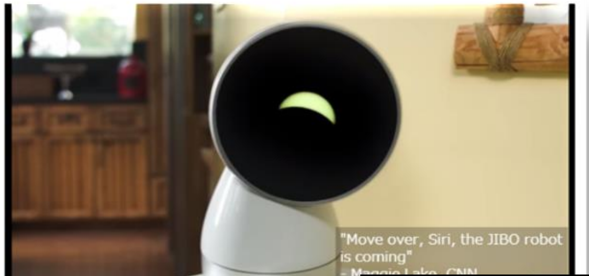
The Family Dog

By Zackary Canepari, Drea Cooper

We have to look deeper to see this connection.

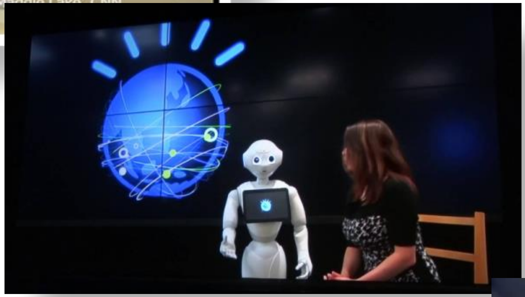
00:44 08:28 HD

소셜 로봇의 사용 확대는 인간과 로봇의 새로운 관계 설정을 필요로 한다



Jibo

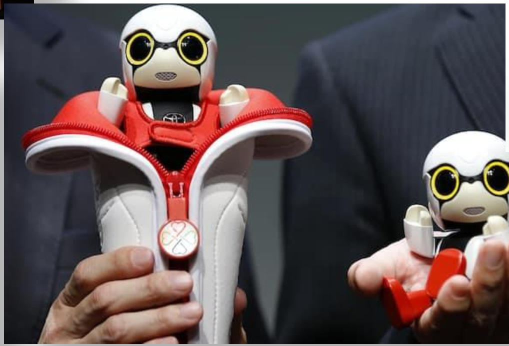
Buddy



Softbank Pepper

KURI

LG Hub



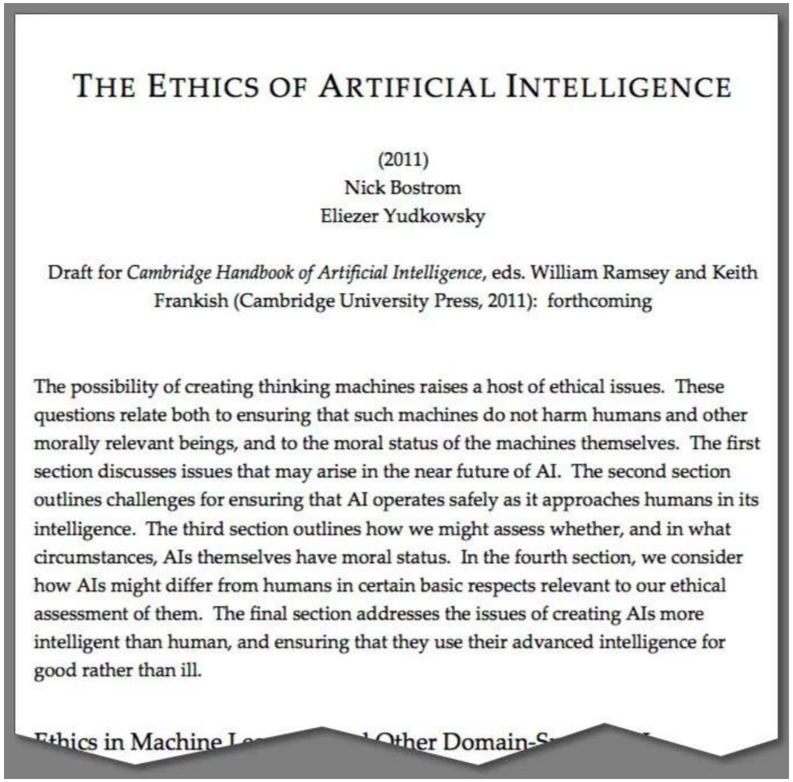
Toyota Kirobo

인공 지능 윤리의 출발점

- 로봇공학자의 전문가적 윤리
- 로봇 안에 프로그램된 '모럴 코드'(moral code)
- 로봇에 의해 윤리적 추론이 이루어질 수 있는 자기 인식 능력을 의미하는 로봇 윤리

+

- 사용자 윤리



THE ETHICS OF ARTIFICIAL INTELLIGENCE

(2011)
Nick Bostrom
Eliezer Yudkowsky

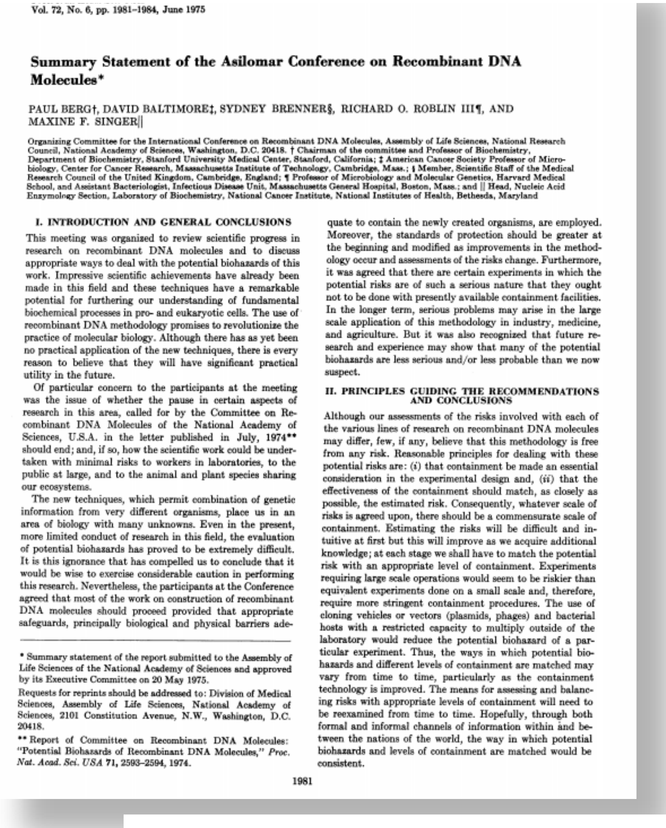
Draft for *Cambridge Handbook of Artificial Intelligence*, eds. William Ramsey and Keith Frankish (Cambridge University Press, 2011): forthcoming

The possibility of creating thinking machines raises a host of ethical issues. These questions relate both to ensuring that such machines do not harm humans and other morally relevant beings, and to the moral status of the machines themselves. The first section discusses issues that may arise in the near future of AI. The second section outlines challenges for ensuring that AI operates safely as it approaches humans in its intelligence. The third section outlines how we might assess whether, and in what circumstances, AIs themselves have moral status. In the fourth section, we consider how AIs might differ from humans in certain basic respects relevant to our ethical assessment of them. The final section addresses the issues of creating AIs more intelligent than human, and ensuring that they use their advanced intelligence for good rather than ill.

Ethics in Machine L Other Domain-S

AI 연구자의 윤리 기준

Summary Statement of the Asilomar Conference on Recombinant DNA in 1975



Vol. 72, No. 6, pp. 1981–1984, June 1975

Summary Statement of the Asilomar Conference on Recombinant DNA Molecules*

PAUL BERG†, DAVID BALTIMORE‡, SYDNEY BRENNER§, RICHARD O. ROBLIN III¶, AND MAXINE F. SINGER||

Organizing Committee for the International Conference on Recombinant DNA Molecules, Assembly of Life Sciences, National Research Council, National Academy of Sciences, Washington, D.C. 20418. † Chairman of the committee and Professor of Biochemistry, Department of Biochemistry, Stanford University Medical Center, Stanford, California; ‡ American Cancer Society Professor of Microbiology, Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Mass.; § Member, Scientific Staff of the Medical Research Council of the United Kingdom, Cambridge, England; ¶ Professor of Microbiology and Molecular Genetics, Harvard Medical School, and Assistant Bacteriologist, Infectious Disease Unit, Massachusetts General Hospital, Boston, Mass.; and || Head, Nucleic Acid Enzymology Section, Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, Maryland

I. INTRODUCTION AND GENERAL CONCLUSIONS

This meeting was organized to review scientific progress in research on recombinant DNA molecules and to discuss appropriate ways to deal with the potential biohazards of this work. Impressive scientific achievements have already been made in this field and these techniques have a remarkable potential for furthering our understanding of fundamental biochemical processes in pro- and eukaryotic cells. The use of recombinant DNA methodology promises to revolutionize the practice of molecular biology. Although there has as yet been no practical application of the new techniques, there is every reason to believe that they will have significant practical utility in the future.

Of particular concern to the participants at the meeting was the issue of whether the pause in certain aspects of research in this area, called for by the Committee on Recombinant DNA Molecules of the National Academy of Sciences, U.S.A. in the letter published in July, 1974** should end; and, if so, how the scientific work could be undertaken with minimal risks to workers in laboratories, to the public at large, and to the animal and plant species sharing our ecosystems.

The new techniques, which permit combination of genetic information from very different organisms, place us in an area of biology with many unknowns. Even in the present, more limited conduct of research in this field, the evaluation of potential biohazards has proved to be extremely difficult. It is this ignorance that has compelled us to conclude that it would be wise to exercise considerable caution in performing this research. Nevertheless, the participants at the Conference agreed that most of the work on construction of recombinant DNA molecules should proceed provided that appropriate safeguards, principally biological and physical barriers adequate to contain the newly created organisms, are employed. Moreover, the standards of protection should be greater at the beginning and modified as improvements in the methodology occur and assessments of the risks change. Furthermore, it was agreed that there are certain experiments in which the potential risks are of such a serious nature that they ought not to be done with presently available containment facilities. In the longer term, serious problems may arise in the large scale application of this methodology in industry, medicine, and agriculture. But it was also recognized that future research and experience may show that many of the potential biohazards are less serious and/or less probable than we now suspect.

II. PRINCIPLES GUIDING THE RECOMMENDATIONS AND CONCLUSIONS

Although our assessments of the risks involved with each of the various lines of research on recombinant DNA molecules may differ, few, if any, believe that this methodology is free from any risk. Reasonable principles for dealing with these potential risks are: (*i*) that containment be made an essential consideration in the experimental design and, (*ii*) that the effectiveness of the containment should match, as closely as possible, the estimated risk. Consequently, whatever scale of risks is agreed upon, there should be a commensurate scale of containment. Estimating the risks will be difficult and intuitive at first but this will improve as we acquire additional knowledge; at each stage we shall have to match the potential risk with an appropriate level of containment. Experiments requiring large scale operations would seem to be riskier than equivalent experiments done on a small scale and, therefore, require more stringent containment procedures. The use of cloning vehicles or vectors (plasmids, phages) and bacterial hosts with a restricted capacity to multiply outside of the laboratory would reduce the potential biohazard of a particular experiment. Thus, the ways in which potential biohazards and different levels of containment are matched may vary from time to time, particularly as the containment technology is improved. The means for assessing and balancing risks with appropriate levels of containment will need to be reexamined from time to time. Hopefully, through both formal and informal channels of information within and between the nations of the world, the way in which potential biohazards and levels of containment are matched would be consistent.

* Summary statement of the report submitted to the Assembly of Life Sciences of the National Academy of Sciences and approved by its Executive Committee on 20 May 1975.

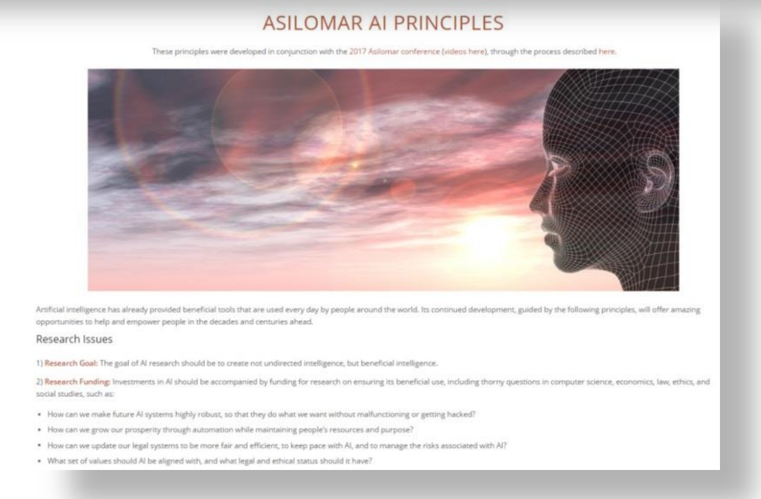
Requests for reprints should be addressed to: Division of Medical Sciences, Assembly of Life Sciences, National Academy of Sciences, 2101 Constitution Avenue, N.W., Washington, D.C. 20418.

** Report of Committee on Recombinant DNA Molecules: "Potential Biohazards of Recombinant DNA Molecules," *Proc. Nat. Acad. Sci. USA* 71, 2593–2594, 1974.

1981

NIH-RAC formed as a result

BENEFICIAL AI 2017



ASILOMAR AI PRINCIPLES

These principles were developed in conjunction with the 2017 Asilomar conference (videos here), through the process described here.

Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead.

Research Issues

1) Research Goal: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.

2) Research Funding: Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as:

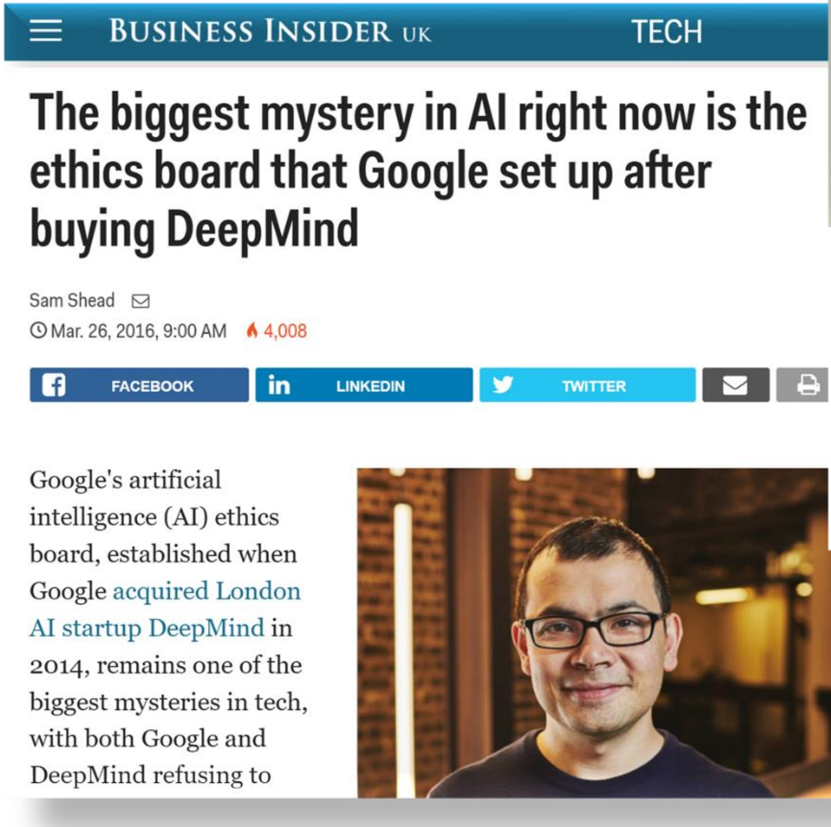
- How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
- How can we grow our prosperity through automation while maintaining people's resources and purpose?
- How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?
- What set of values should AI be aligned with, and what legal and ethical status should it have?

[출처: Future of Life Institute]

Asilomar AI Ethics and Values

- 6) **Safety:** AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.7) **Failure Transparency:** If an AI system causes harm, it should be possible to ascertain why.
- 8) **Judicial Transparency:** Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
- 9) **Responsibility:** Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
- 10) **Value Alignment:** Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
- 11) **Human Values:** AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
- 12) **Personal Privacy:** People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
- 13) **Liberty and Privacy:** The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
- 14) **Shared Benefit:** AI technologies should benefit and empower as many people as possible.
- 15) **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
- 16) **Human Control:** Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
- 17) **Non-subversion:** The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
- 18) **AI Arms Race:** An arms race in lethal autonomous weapons should be avoided.

기업 윤리 위원회의 투명성은 어디까지?



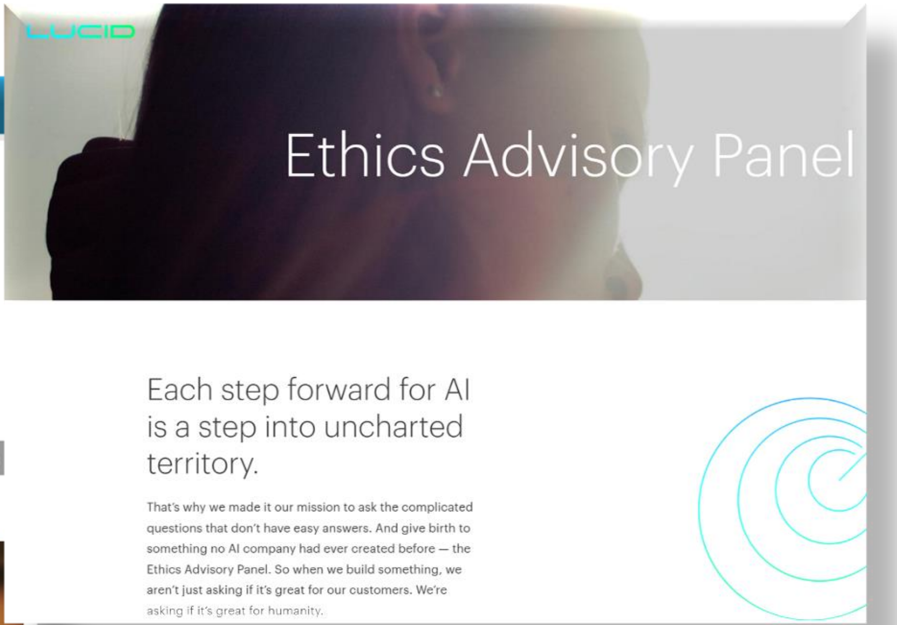
BUSINESS INSIDER UK TECH

The biggest mystery in AI right now is the ethics board that Google set up after buying DeepMind

Sam Shead

Mar. 26, 2016, 9:00 AM 4,008

Google's artificial intelligence (AI) ethics board, established when Google acquired London AI startup DeepMind in 2014, remains one of the biggest mysteries in tech, with both Google and DeepMind refusing to



Ethics Advisory Panel

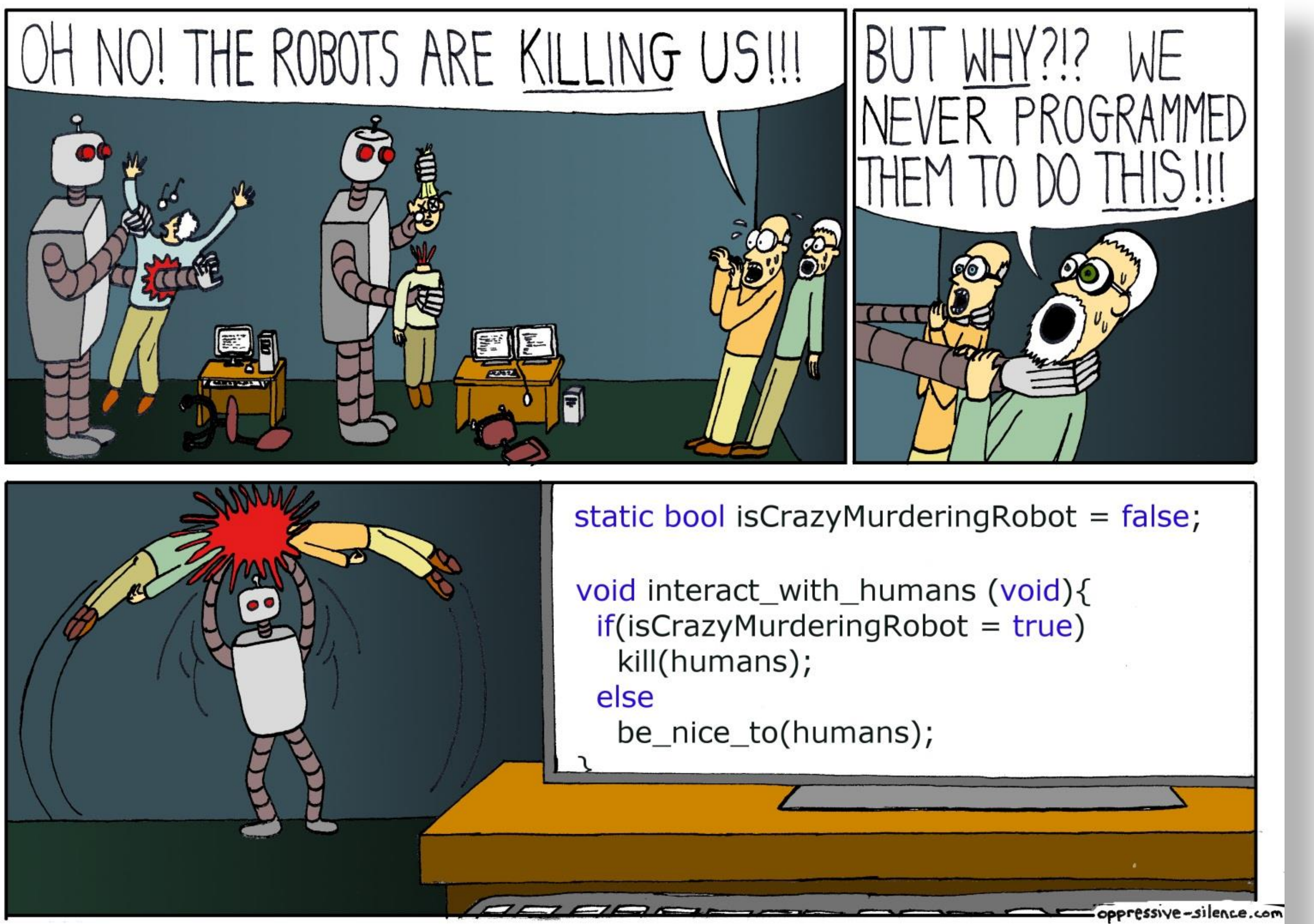
Each step forward for AI is a step into uncharted territory.

That's why we made it our mission to ask the complicated questions that don't have easy answers. And give birth to something no AI company had ever created before — the Ethics Advisory Panel. So when we build something, we aren't just asking if it's great for our customers. We're asking if it's great for humanity.

1. Ethics Isn't Just About Legal Risk
2. Internal vs. External Advisors: Pros And Cons
3. More than Lip-Service About Ethics

-- Patrick Lin and Evan Selinger

모랄 코드는 구현이 가능할까?

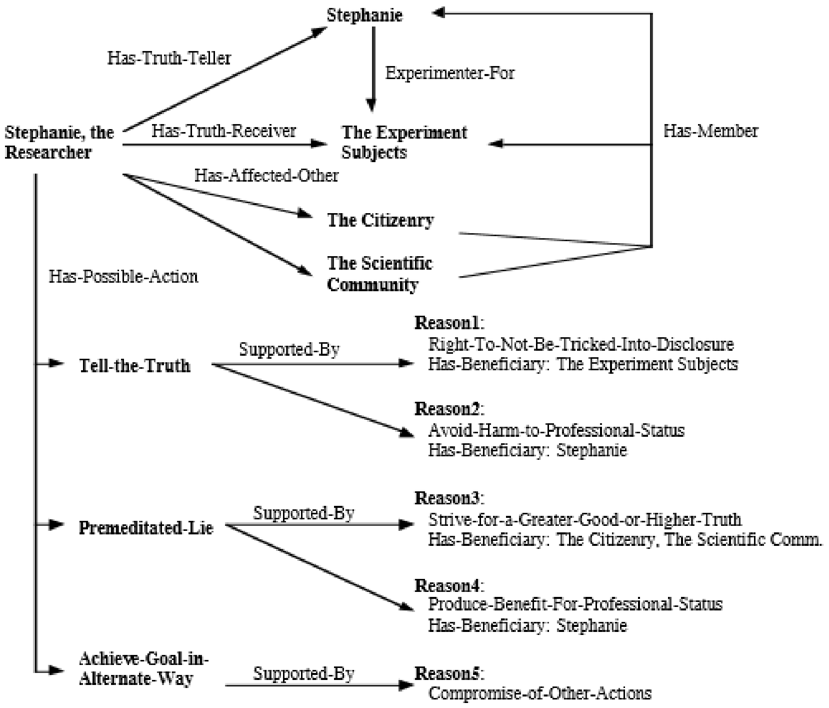

OH NO! THE ROBOTS ARE KILLING US!!!
BUT WHY?!? WE NEVER PROGRAMMED THEM TO DO THIS!!!
oppressive-silence.com

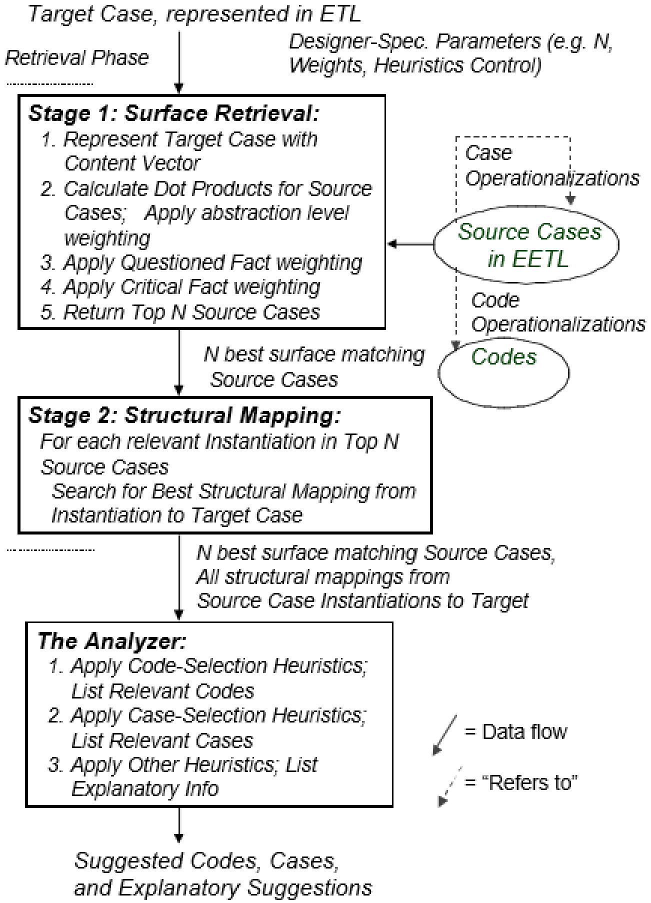
```
static bool isCrazyMurderingRobot = false;

void interact_with_humans (void){
  if(isCrazyMurderingRobot = true)
    kill(humans);
  else
    be_nice_to(humans);
}
```

[출처: Reddit]

Early Computational Models of Ethical Reasoning: Truth-Teller and SIROCCO


Stephanie
Has-Truth-Teller
Experimenter-For
Stephanie, the Researcher
Has-Truth-Receiver
The Experiment Subjects
Has-Member
Has-Affected-Other
The Citizenry
The Scientific Community
Has-Possible-Action
Tell-the-Truth
Supported-By
Reason1:
Right-To-Not-Be-Tricked-Into-Disclosure
Has-Beneficiary: The Experiment Subjects
Reason2:
Avoid-Harm-to-Professional-Status
Has-Beneficiary: Stephanie
Premeditated-Lie
Supported-By
Reason3:
Strive-for-a-Greater-Good-or-Higher-Truth
Has-Beneficiary: The Citizenry, The Scientific Comm.
Reason4:
Produce-Benefit-For-Professional-Status
Has-Beneficiary: Stephanie
Achieve-Goal-in-Alternate-Way
Supported-By
Reason5:
Compromise-of-Other-Actions


Target Case, represented in ETL
Retrieval Phase
Designer-Spec. Parameters (e.g. N, Weights, Heuristics Control)
Case Operationalizations
Source Cases in EETL
Code Operationalizations
Codes
N best surface matching Source Cases
N best surface matching Source Cases, All structural mappings from Source Case Instantiations to Target
= Data flow
= "Refers to"
Suggested Codes, Cases, and Explanatory Suggestions

Stage 1: Surface Retrieval:

1. *Represent Target Case with Content Vector*
2. *Calculate Dot Products for Source Cases; Apply abstraction level weighting*
3. *Apply Questioned Fact weighting*
4. *Apply Critical Fact weighting*
5. *Return Top N Source Cases*

Stage 2: Structural Mapping:

For each relevant Instantiation in Top N Source Cases
Search for Best Structural Mapping from Instantiation to Target Case

The Analyzer:

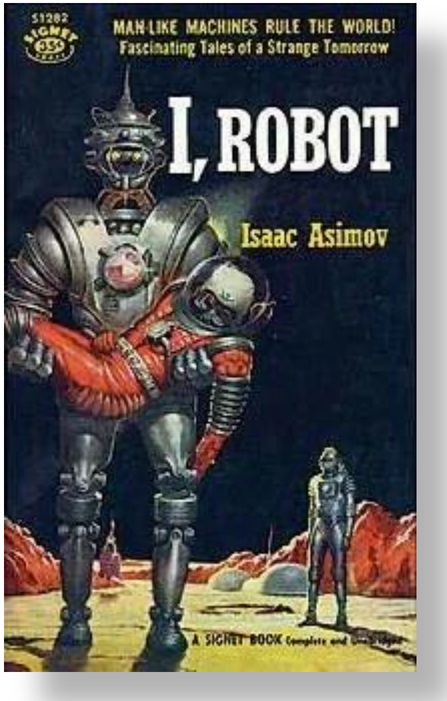
1. *Apply Code-Selection Heuristics; List Relevant Codes*
2. *Apply Case-Selection Heuristics; List Relevant Cases*
3. *Apply Other Heuristics; List Explanatory Info*

Bruce M. McLaren @ CMU

Institute of Social Computing
소셜컴퓨팅연구소

아시모프의 로봇 3원칙의 문제

1. **로봇은 인간에게 해를 가하거나 해를 당하는 상황에서 무시하면 안 된다**
   - '인간'을 어느 개념으로 정의할 것인지부터 쉬운 문제가 아니다. 인류는 한동안 다른 인종을 인간과 다르다고 분류한 적도 있고, 앞으로는 생물학적 특징만으로 인간을 정의하기 어려워질 수 있기 때문이다. 로봇에 이를 주입하는 것이 쉬운 일은 아니다.
   - '해를 가한다'는 것을 판단하려면 우선 그 행동의 결과가 누구에게 해가 될 수 있는지를 알아야 한다. 하지만 모든 상황을 정확히 판단한다는 것은 불가능한 일이다. 주변의 인간에게는 해를 가하지 않아도 지구 어딘가에 있는 다른 인간에게 해를 끼칠 수도 있는데 이를 계산할 방안이 없다.
2. **로봇은 1원칙에 어긋나지 않는 한, 인간의 명령에 복종해야 한다**
   - 인간의 발언 중 어디까지가 명령임을 알아내는 것 자체가 쉬운 일이 아니다. 제1원칙과 제2원칙을 따라서 행동한 것이 특정인을 구했지만, 그 결과로 인류에게 엄청난 파국을 일으킨다면 어떻게 할 것인가?
3. **로봇은 1, 2원칙에 어긋나지 않는 한, 자신을 지켜야 한다.**

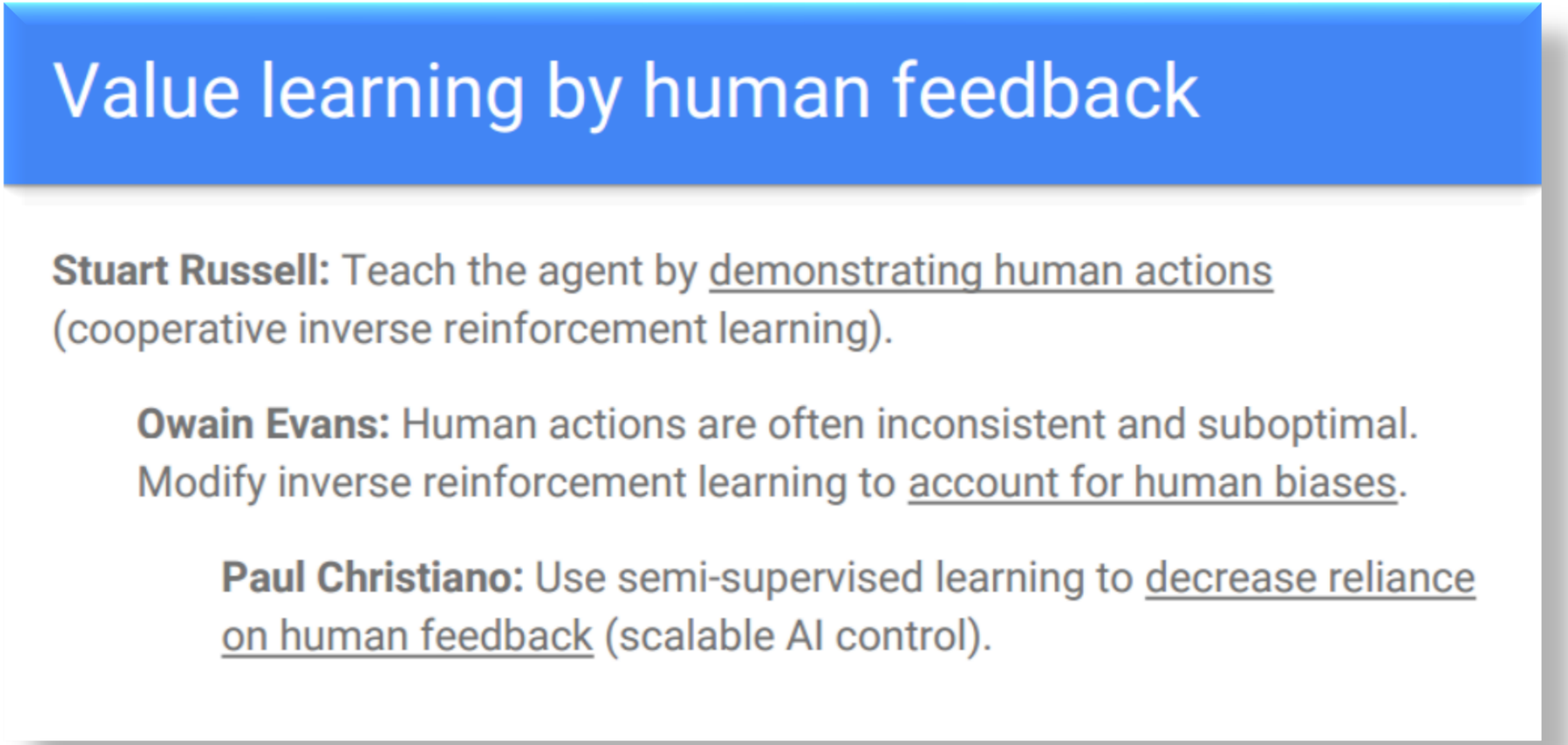

S1282
SIGNET
35¢
MAN-LIKE MACHINES RULE THE WORLD!
Fascinating Tales of a Strange Tomorrow
I, ROBOT
Isaac Asimov
A SIGNET BOOK Complete and Unabridged

- 0원칙 (1985): **로봇은 인류에게 해를 가하거나, 또는 해를 당하는 상황을 무시해서는 안 된다.**

➔ 의무론적이고 하향식 방법의 문제는 규칙을 완벽하게 따를 때도 여전히 끔찍한 결과를 일으킬 수 있다

Value learning

Question: How to specify complicated human values and ethics to AI systems?



Value learning by human feedback

Stuart Russell: Teach the agent by demonstrating human actions (cooperative inverse reinforcement learning).

Owain Evans: Human actions are often inconsistent and suboptimal. Modify inverse reinforcement learning to account for human biases.

Paul Christiano: Use semi-supervised learning to decrease reliance on human feedback (scalable AI control).

Viktoriya Krakovna, FLI / DeepMind @BAI2017

Value Learning

Value learning by building in morality

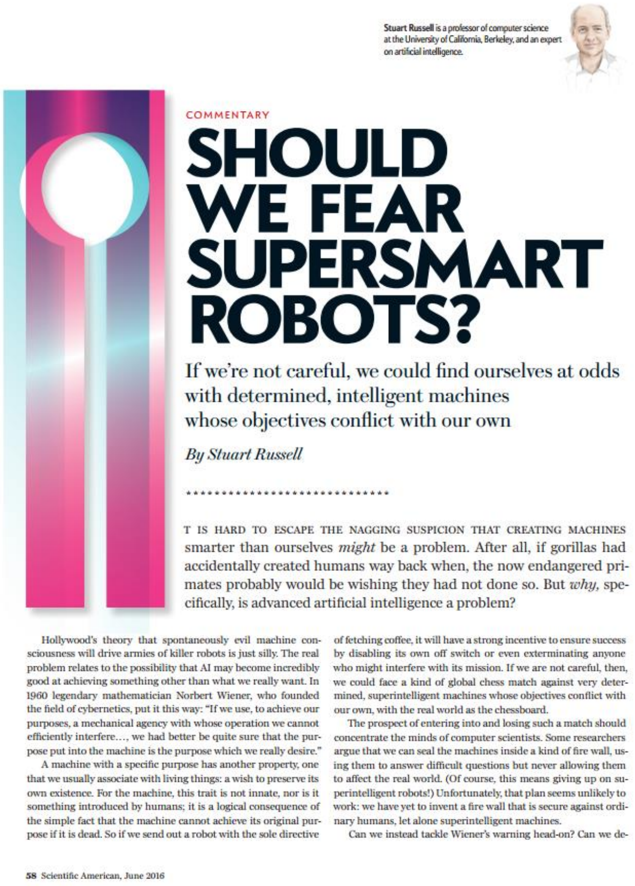
Francesca Rossi: Specify ethical laws through constraints.

Vincent Conitzer: Find patterns in human ethical decisions, and build those features into AI systems.

Adrian Weller: Can we make human moral concepts more precise and consistent?

Viktoriya Krakovna, FLI / DeepMind @BAI2017

상향식 접근: Learn by Observation
"How to Stop Your Robot Cooking Your Cat"



Stuart Russell is a professor of computer science at the University of California, Berkeley, and an expert on artificial intelligence.

COMMENTARY

SHOULD WE FEAR SUPERSMART ROBOTS?

If we're not careful, we could find ourselves at odds with determined, intelligent machines whose objectives conflict with our own

By Stuart Russell

IT IS HARD TO ESCAPE THE NAGGING SUSPICION THAT CREATING MACHINES smarter than ourselves *might* be a problem. After all, if gorillas had accidentally created humans way back when, the now endangered primates probably would be wishing they had not done so. But *why*, specifically, is advanced artificial intelligence a problem?

Hollywood's theory that spontaneously evil machine consciousness will drive armies of killer robots is just silly. The real problem relates to the possibility that AI may become incredibly good at achieving something other than what we really want. In 1960 legendary mathematician Norbert Wiener, who founded the field of cybernetics, put it this way: "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere…, we had better be quite sure that the purpose put into the machine is the purpose which we really desire."

A machine with a specific purpose has another property, one that we usually associate with living things: a wish to preserve its own existence. For the machine, this trait is not innate, nor is it something introduced by humans; it is a logical consequence of the simple fact that the machine cannot achieve its original purpose if it is dead. So if we send out a robot with the sole directive of fetching coffee, it will have a strong incentive to ensure success by disabling its own off switch or even exterminating anyone who might interfere with its mission. If we are not careful, then, we could face a kind of global chess match against very determined, superintelligent machines whose objectives conflict with our own, with the real world as the chessboard.

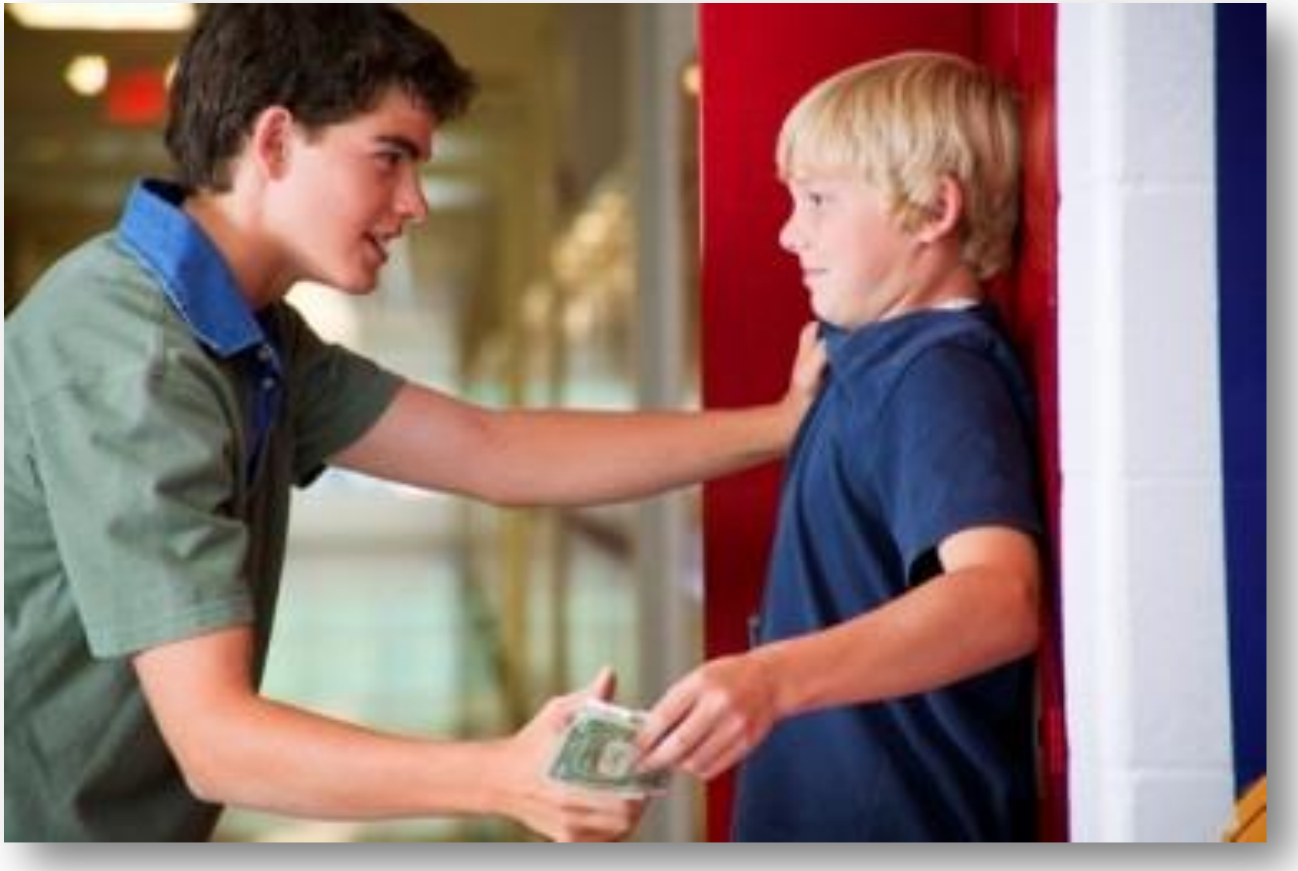
The prospect of entering into and losing such a match should concentrate the minds of computer scientists. Some researchers argue that we can seal the machines inside a kind of fire wall, using them to answer difficult questions but never allowing them to affect the real world. (Of course, this means giving up on superintelligent robots!) Unfortunately, that plan seems unlikely to work: we have yet to invent a fire wall that is secure against ordinary humans, let alone superintelligent machines.

Can we instead tackle Wiener's warning head-on? Can we de-

58 Scientific American, June 2016

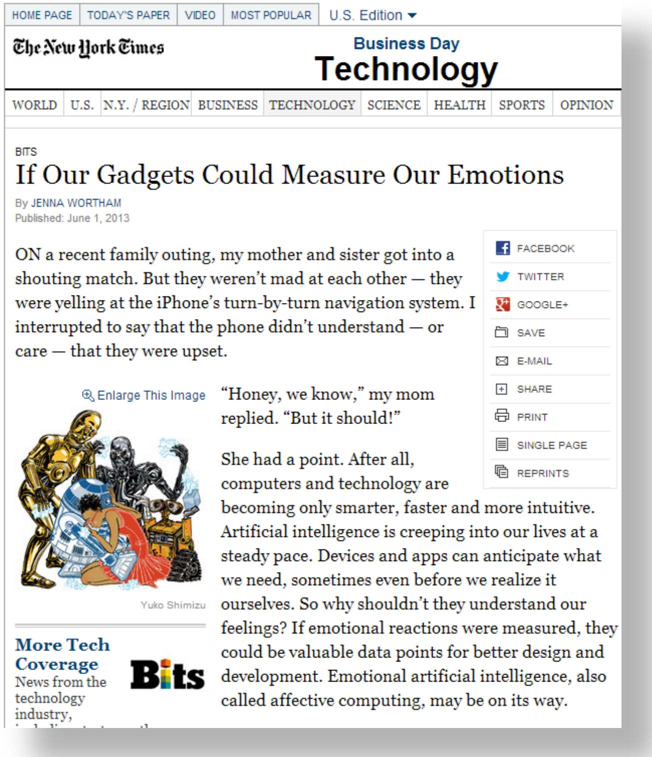
- The machine's purpose must be to maximize the realization of human values. In particular, it has no purpose of its own and no innate desire to protect itself.
- The machine must be initially uncertain about what those human values are. The machine may learn more about human values as it goes along, of course, but it may never achieve complete certainty.
- The machine must be able to learn about human values by observing the choices that we humans make.
  - Inverse reinforcement learning (IRL), concerned with learning the values of some by observing its behavior. By watching a typical human's morning routine, the robot learns about the value of coffee to humans.
- 상향식 접근은 인간 행동의 목적과 결과, 영향, 행동이 윤리적 기반을 갖는 것임을 판단하는 능력을 갖추고, 이를 다시 내부의 코드로 만들어가야 하는 어려움을 갖고 있다.
- 인간 가치를 표현하고 이를 각 인간의 배경에 따라 윤리, 법, 도덕으로 인지할 수 있는 능력을 포함해야 하므로 아직 이런 방향의 연구는 초기 단계라고 생각한다. 더군다나 이런 인간 행동이 윤리적 의미를 갖게 될 때에는 다양한 인간 감성 표현과 감정을 이해해야 하는데, 이는 아직 우리가 알고 있는 인공지능 기술에서도 매우 어려운 분야이다.

상황 인지의 어려움



감정을 어떻게 인식할 것인가?
Affective Computing

- The field of study concerned with understanding, recognizing, and utilizing human emotions and other affective phenomena in the design of technological systems - IEEE



HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition ▾

The New York Times

Business Day
Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

BITS

If Our Gadgets Could Measure Our Emotions

By JENNA WORTHAM
Published: June 1, 2013

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT
SINGLE PAGE
REPRINTS

ON a recent family outing, my mother and sister got into a shouting match. But they weren't mad at each other — they were yelling at the iPhone's turn-by-turn navigation system. I interrupted to say that the phone didn't understand — or care — that they were upset.

Enlarge This Image

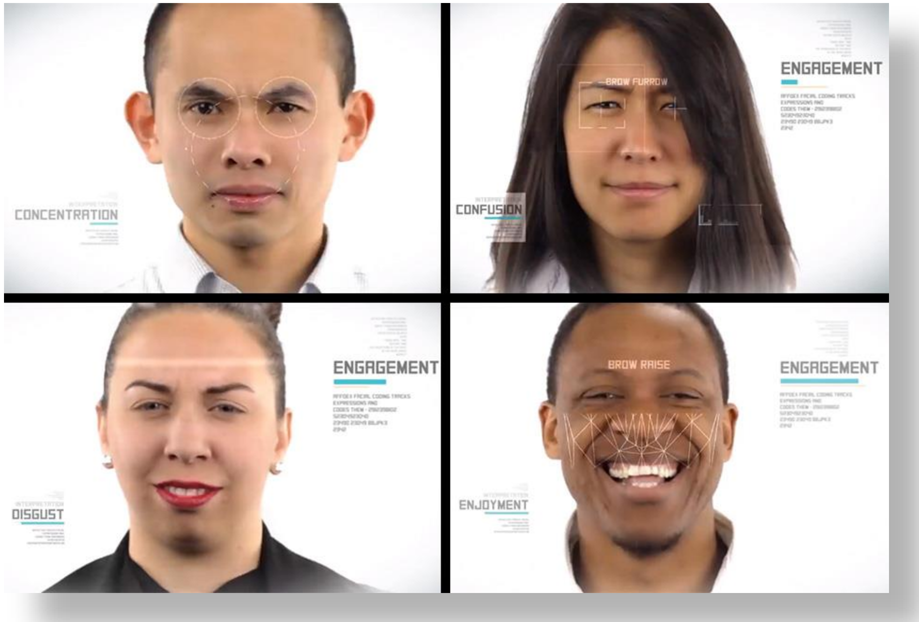
Yuko Shimizu

"Honey, we know," my mom replied. "But it should!"

She had a point. After all, computers and technology are becoming only smarter, faster and more intuitive. Artificial intelligence is creeping into our lives at a steady pace. Devices and apps can anticipate what we need, sometimes even before we realize it ourselves. So why shouldn't they understand our feelings? If emotional reactions were measured, they could be valuable data points for better design and development. Emotional artificial intelligence, also called affective computing, may be on its way.

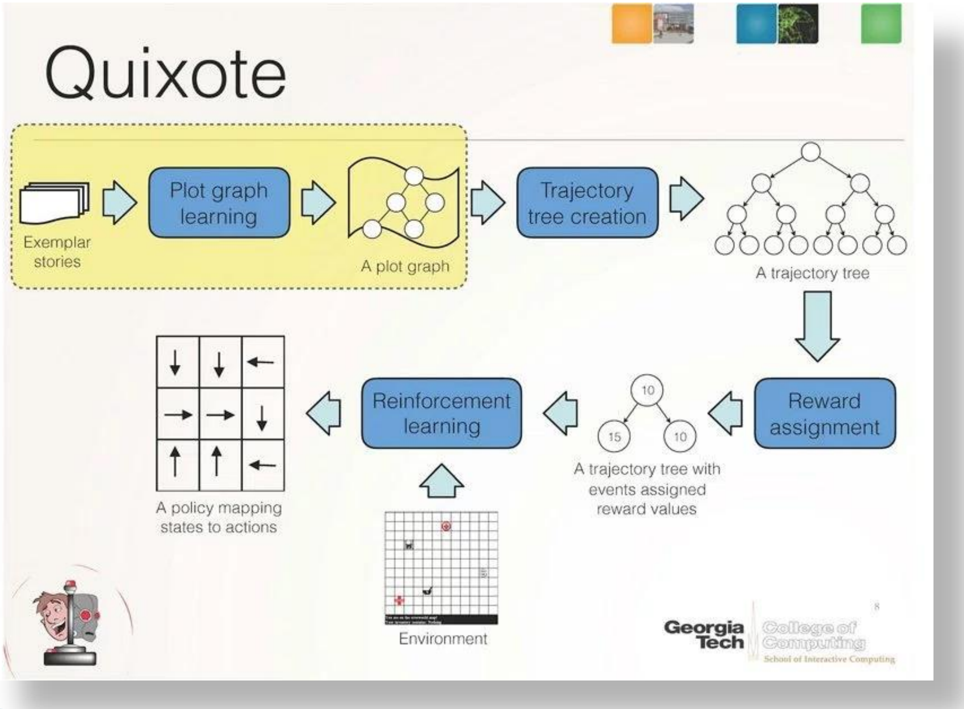
More Tech Coverage
News from the technology industry,

Bits

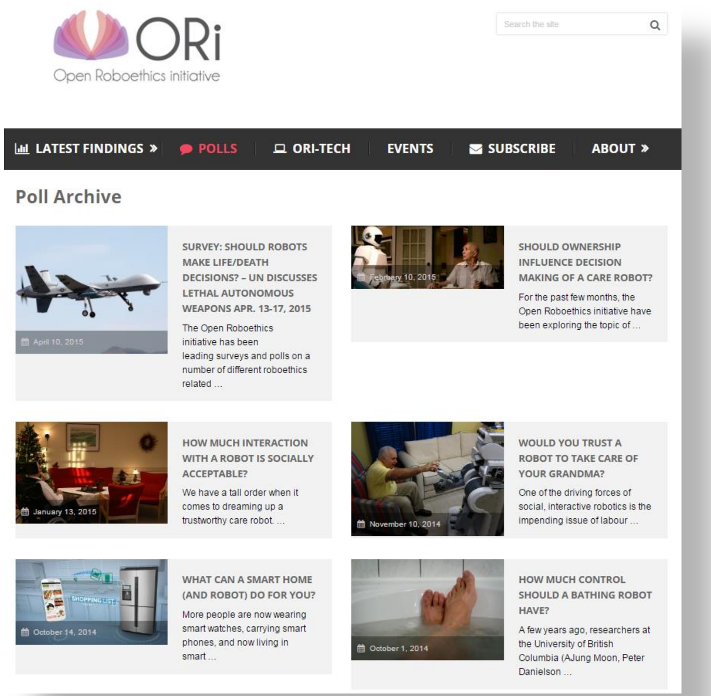


Using Stories to Teach Human Values to Artificial Agents

- Mark Riedl and Brent Harrison. 2016. Using Stories to Teach Human Values to Artificial Agents. In Proceedings of the 2nd International Workshop on AI, Ethics and Society.
- 로봇이나 인공 행위자(에이전트)가 이야기를 읽고, 각 사건의 바람직한 결과를 학습해, 인간 사회에서 성공적인 행동을 이해하도록 훈련하는 시스템


Quixote

사람들의 참여에 의한 윤리 판단 데이터 수집



Poll Archive

April 10, 2015

SURVEY: SHOULD ROBOTS MAKE LIFE/DEATH DECISIONS? – UN DISCUSSES LETHAL AUTONOMOUS WEAPONS APR. 13-17, 2015

The Open Roboethics initiative has been leading surveys and polls on a number of different roboethics related ...

February 10, 2015

SHOULD OWNERSHIP INFLUENCE DECISION MAKING OF A CARE ROBOT?

For the past few months, the Open Roboethics initiative have been exploring the topic of ...

January 13, 2015

HOW MUCH INTERACTION WITH A ROBOT IS SOCIALLY ACCEPTABLE?

We have a tall order when it comes to dreaming up a trustworthy care robot. ...

November 10, 2014

WOULD YOU TRUST A ROBOT TO TAKE CARE OF YOUR GRANDMA?

One of the driving forces of social, interactive robotics is the impending issue of labour ...

October 14, 2014

WHAT CAN A SMART HOME (AND ROBOT) DO FOR YOU?

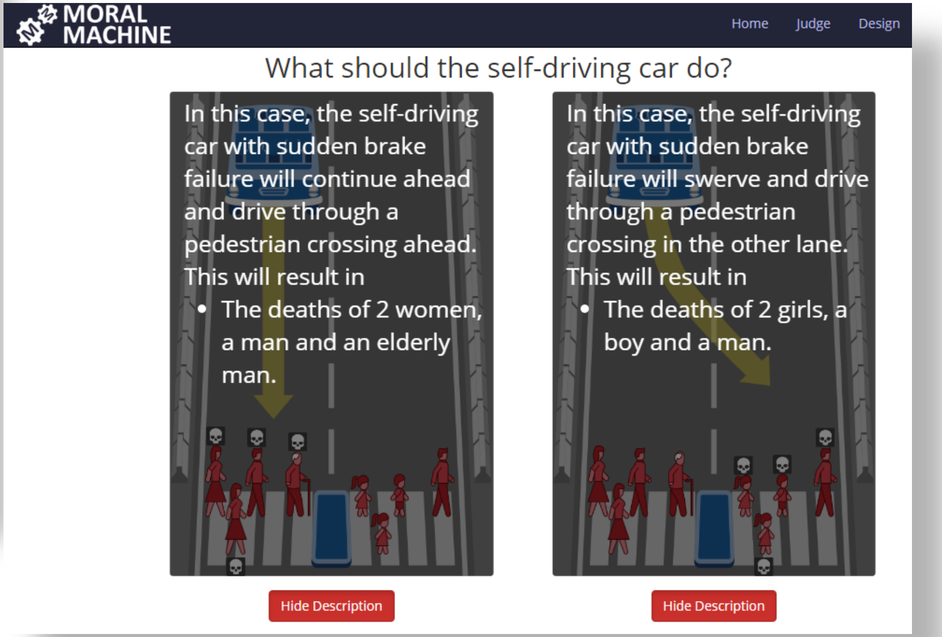
More people are now wearing smart watches, carrying smart phones, and now living in smart ...

October 1, 2014

HOW MUCH CONTROL SHOULD A BATHING ROBOT HAVE?

A few years ago, researchers at the University of British Columbia (AJung Moon, Peter Danielson ...

Providing a platform for 1) building a crowd-sourced picture of human opinion on how machines should make decisions when faced with moral dilemmas, and 2) crowd-sourcing assembly and discussion of potential scenarios of moral consequence



What should the self-driving car do?

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

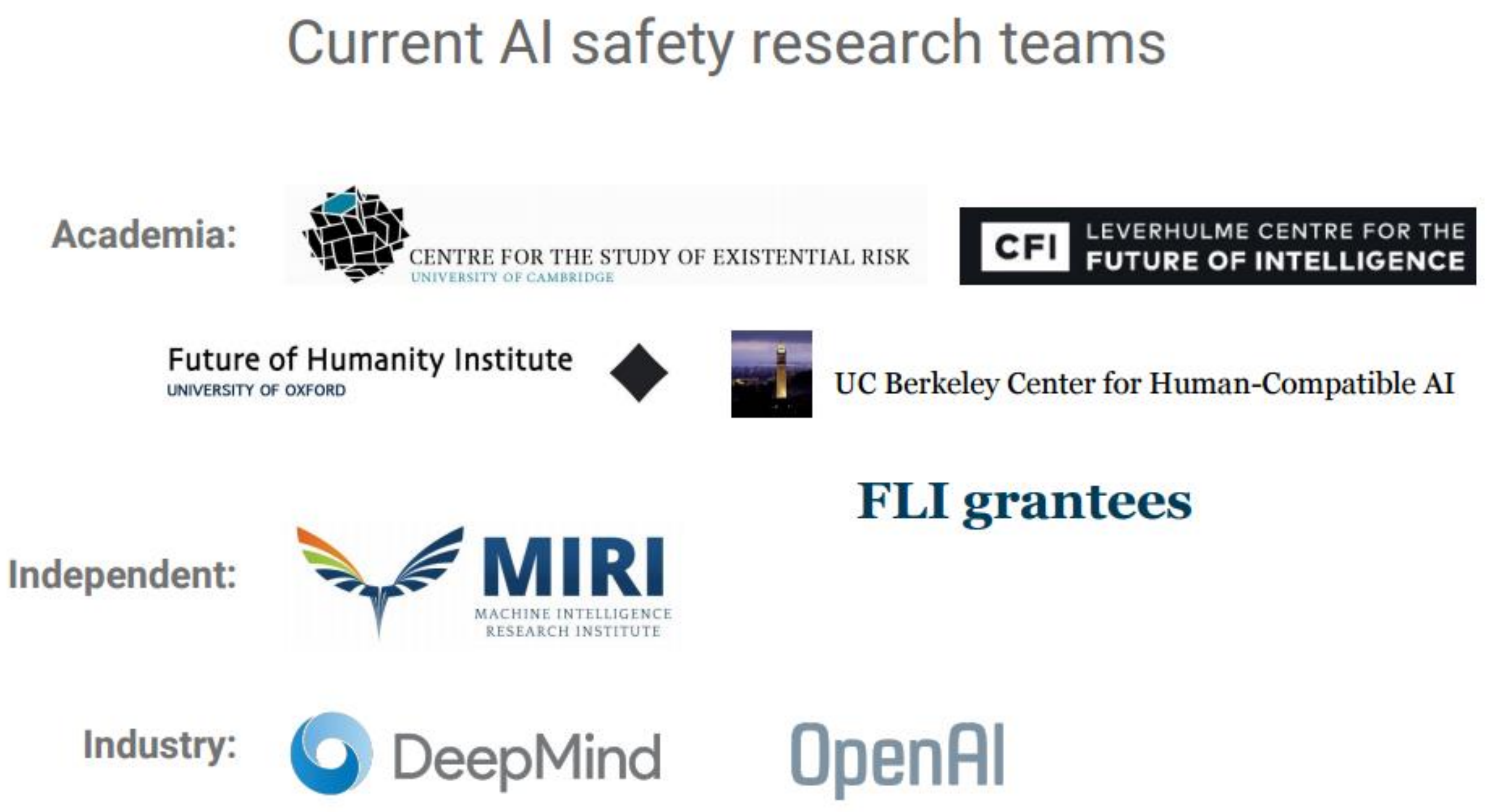
- The deaths of 2 women, a man and an elderly man.

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in

- The deaths of 2 girls, a boy and a man.

http://moralmachine.mit.edu/

AI Safety Research Teams



Current AI safety research teams

Academia:

CENTRE FOR THE STUDY OF EXISTENTIAL RISK
UNIVERSITY OF CAMBRIDGE

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Future of Humanity Institute
UNIVERSITY OF OXFORD

UC Berkeley Center for Human-Compatible AI

FLI grantees

Independent:

MIRI
MACHINE INTELLIGENCE RESEARCH INSTITUTE

Industry:

DeepMind

OpenAI

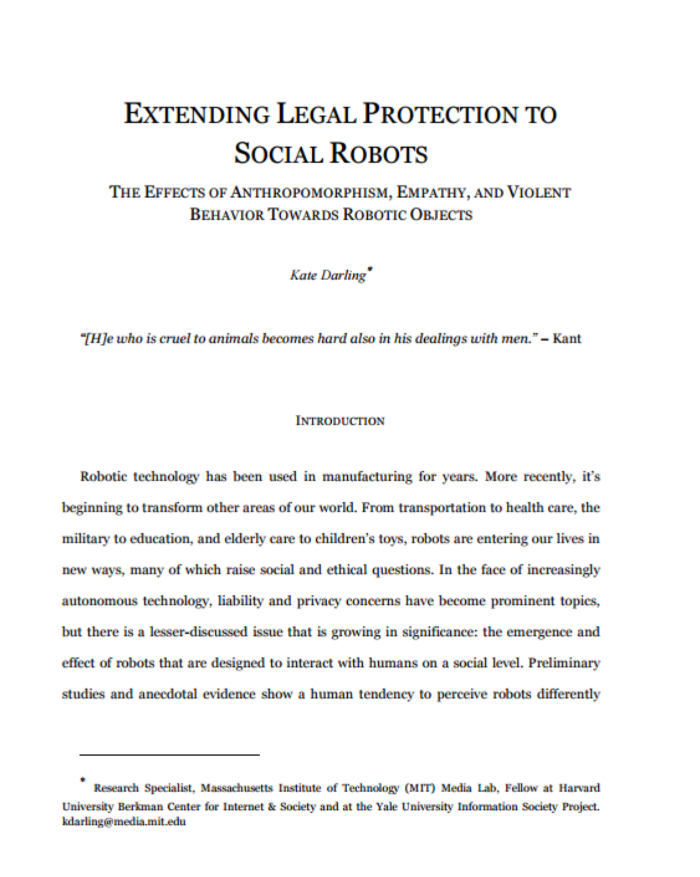
Challenges and Further Researches

- Transparency
- Explainability
- Computational model of ethics
- How Do We Align Artificial Intelligence with Human Values?
  - Understanding what "we" want
  - Humanity do not agree on common values, and even parts we do agree on change with time
- Collaboration between AI researchers and Ethicists

감사합니다

(Meet me at 페이스북: facebook.com/stevehan)

소셜 로봇에 대한 법적 보호와 안전성 요구 문제



Extending Legal Protection to Social Robots

The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects

Kate Darling[*]

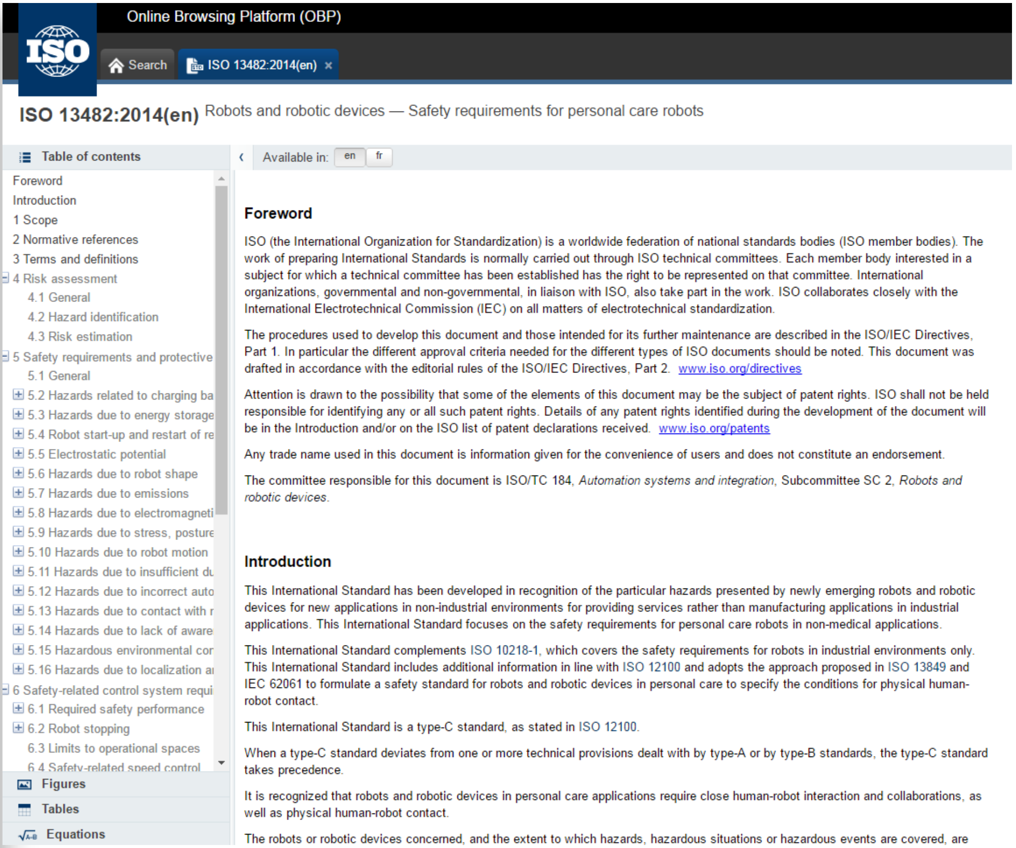
"[H]e who is cruel to animals becomes hard also in his dealings with men." – Kant

Introduction

Robotic technology has been used in manufacturing for years. More recently, it's beginning to transform other areas of our world. From transportation to health care, the military to education, and elderly care to children's toys, robots are entering our lives in new ways, many of which raise social and ethical questions. In the face of increasingly autonomous technology, liability and privacy concerns have become prominent topics, but there is a lesser-discussed issue that is growing in significance: the emergence and effect of robots that are designed to interact with humans on a social level. Preliminary studies and anecdotal evidence show a human tendency to perceive robots differently

[*] Research Specialist, Massachusetts Institute of Technology (MIT) Media Lab, Fellow at Harvard University Berkman Center for Internet & Society and at the Yale University Information Society Project. kdarling@media.mit.edu

[출처: Kate Darling, MIT]



ISO 13482:2014(en) Robots and robotic devices — Safety requirements for personal care robots

Table of contents

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2. www.iso.org/directives

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received. www.iso.org/patents

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

The committee responsible for this document is ISO/TC 184, *Automation systems and integration*, Subcommittee SC 2, *Robots and robotic devices*.

Introduction

This International Standard has been developed in recognition of the particular hazards presented by newly emerging robots and robotic devices for new applications in non-industrial environments for providing services rather than manufacturing applications in industrial applications. This International Standard focuses on the safety requirements for personal care robots in non-medical applications.

This International Standard complements ISO 10218-1, which covers the safety requirements for robots in industrial environments only. This International Standard includes additional information in line with ISO 12100 and adopts the approach proposed in ISO 13849 and IEC 62061 to formulate a safety standard for robots and robotic devices in personal care to specify the conditions for physical human-robot contact.

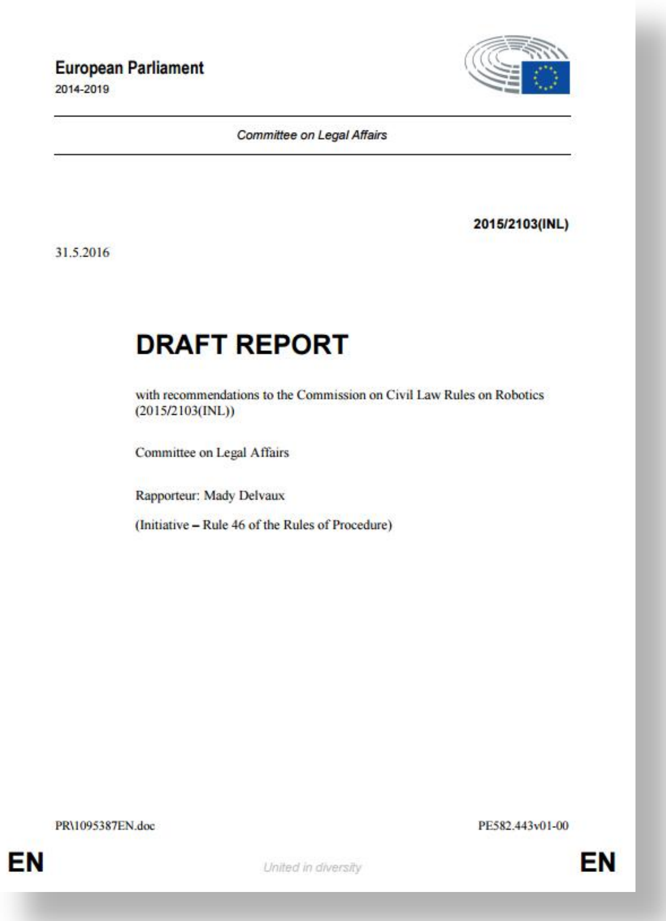
This International Standard is a type-C standard, as stated in ISO 12100.

When a type-C standard deviates from one or more technical provisions dealt with by type-A or by type-B standards, the type-C standard takes precedence.

It is recognized that robots and robotic devices in personal care applications require close human-robot interaction and collaborations, as well as physical human-robot contact.

The robots or robotic devices concerned, and the extent to which hazards, hazardous situations or hazardous events are covered, are

Civil Law Rules on Robotics:
Draft Report of EU Parliament



European Parliament
2014-2019

Committee on Legal Affairs

2015/2103(INL)

31.5.2016

DRAFT REPORT

with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))

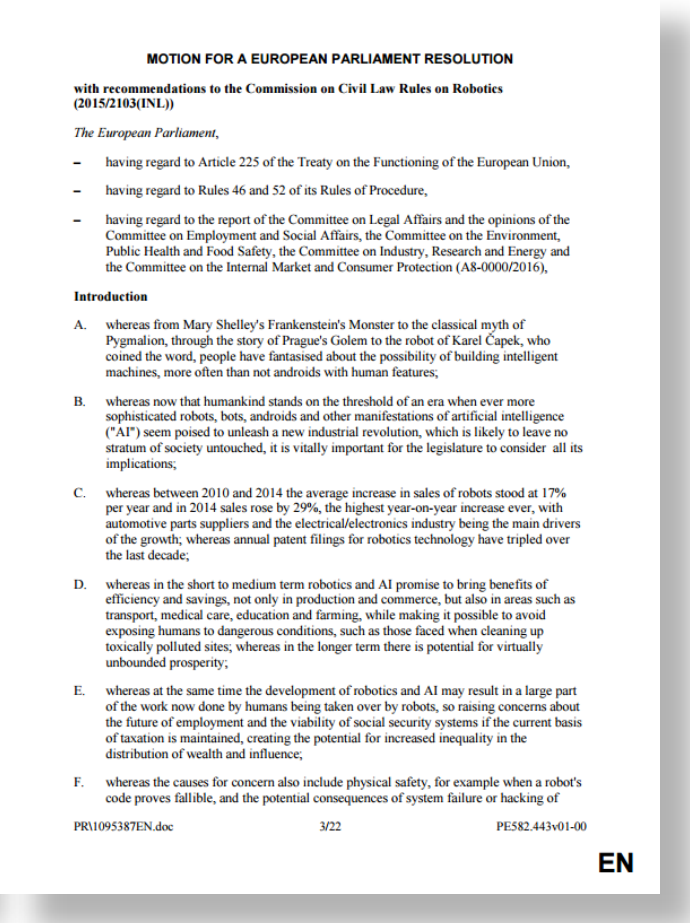
Committee on Legal Affairs

Rapporteur: Mady Delvaux

(Initiative – Rule 46 of the Rules of Procedure)

PR\1095387EN.doc PE582.443v01-00

EN *United in diversity* EN



MOTION FOR A EUROPEAN PARLIAMENT RESOLUTION

with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))

The European Parliament,

– having regard to Article 225 of the Treaty on the Functioning of the European Union,

– having regard to Rules 46 and 52 of its Rules of Procedure,

– having regard to the report of the Committee on Legal Affairs and the opinions of the Committee on Employment and Social Affairs, the Committee on the Environment, Public Health and Food Safety, the Committee on Industry, Research and Energy and the Committee on the Internal Market and Consumer Protection (A8-0000/2016),

Introduction

A. whereas from Mary Shelley's Frankenstein's Monster to the classical myth of Pygmalion, through the story of Prague's Golem to the robot of Karel Čapek, who coined the word, people have fantasised about the possibility of building intelligent machines, more often than not androids with human features;

B. whereas now that humankind stands on the threshold of an era when ever more sophisticated robots, bots, androids and other manifestations of artificial intelligence ("AI") seem poised to unleash a new industrial revolution, which is likely to leave no stratum of society untouched, it is vitally important for the legislature to consider all its implications;

C. whereas between 2010 and 2014 the average increase in sales of robots stood at 17% per year and in 2014 sales rose by 29%, the highest year-on-year increase ever, with automotive parts suppliers and the electrical/electronics industry being the main drivers of the growth; whereas annual patent filings for robotics technology have tripled over the last decade;

D. whereas in the short to medium term robotics and AI promise to bring benefits of efficiency and savings, not only in production and commerce, but also in areas such as transport, medical care, education and farming, while making it possible to avoid exposing humans to dangerous conditions, such as those faced when cleaning up toxically polluted sites; whereas in the longer term there is potential for virtually unbounded prosperity;

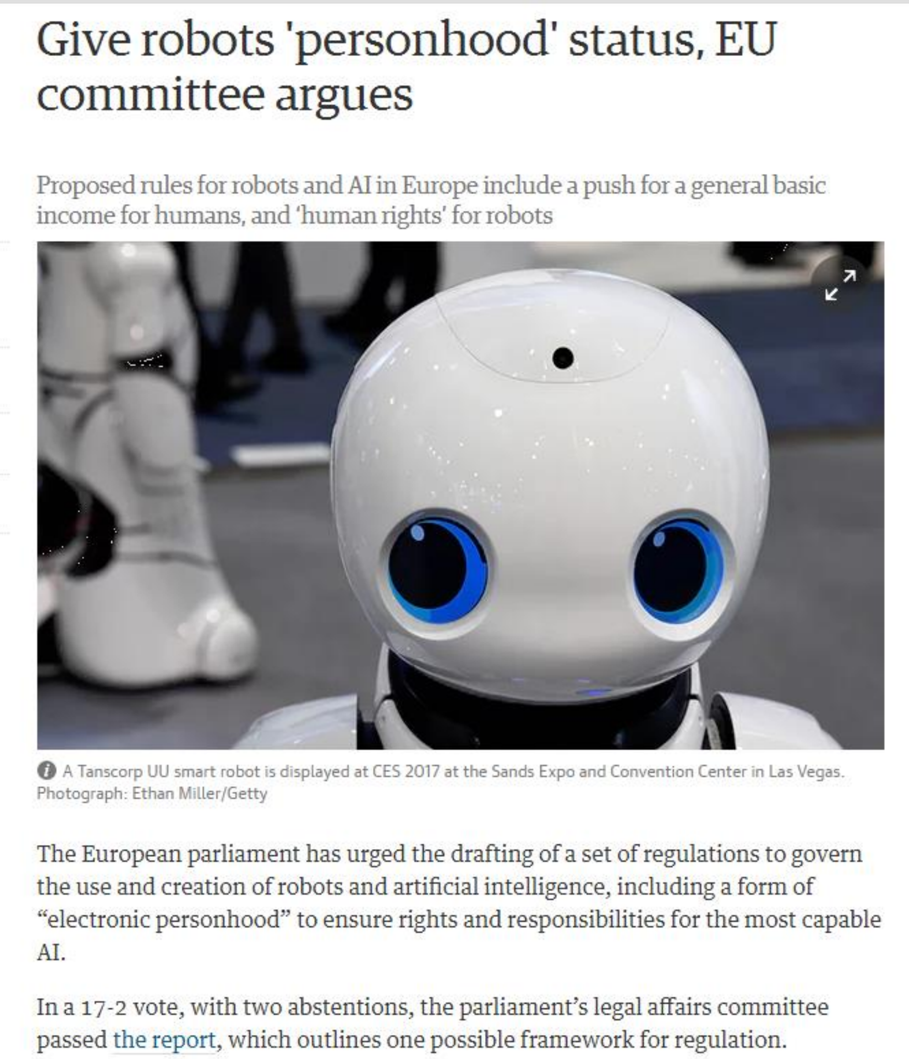
E. whereas at the same time the development of robotics and AI may result in a large part of the work now done by humans being taken over by robots, so raising concerns about the future of employment and the viability of social security systems if the current basis of taxation is maintained, creating the potential for increased inequality in the distribution of wealth and influence;

F. whereas the causes for concern also include physical safety, for example when a robot's code proves fallible, and the potential consequences of system failure or hacking of

PR\1095387EN.doc 3/22 PE582.443v01-00

EN

Electronic Personhood?



Give robots 'personhood' status, EU committee argues

Proposed rules for robots and AI in Europe include a push for a general basic income for humans, and 'human rights' for robots

A Tanscorp UU smart robot is displayed at CES 2017 at the Sands Expo and Convention Center in Las Vegas. Photograph: Ethan Miller/Getty

The European parliament has urged the drafting of a set of regulations to govern the use and creation of robots and artificial intelligence, including a form of "electronic personhood" to ensure rights and responsibilities for the most capable AI.

In a 17-2 vote, with two abstentions, the parliament's legal affairs committee passed the report, which outlines one possible framework for regulation.