Section 6

Ethics and Intelligent Machines

sinead.barton@mu.ie

What are ethics?

- Ethics involves the group and community at the level of **values** – not about **what** can be done, or **how** something is done, but what **should** be done

- Ethics looks at our **responsibility to each other** – to individuals and to the community

- Being ethical is about choosing the morally correct thing even if there is no law to enforce it.

How can ethics be influenced?

- Our sense of ethics is defined by the community
- Many communities have a theological, cultural, or philosophical basis to their civil ethical framework, some have an influence, and some have none
- For example, in Africa Ubuntuism is a type of philosophy that began in the 19th century and still has traditional influences
- This is a complicated space, but clearly this can vary quite a lot from one country to another.
- We are not here to talk about religion, philosophy, or culture but it must be acknowledged that these things can alter how humans can perceive or programme AI.

Remember!!!

- Ethical frameworks vary from region to region.
- What is experienced in the European-North American domain is not the same everywhere else.

How can we apply ethical issues to machines?

Does a machine have a soul?
Can it sin?
Can a machine be un-ethical?

Strong AI vs Weak AI

What is it and how do these approaches affect ethical decisions?

Weak AI vs Strong AI

WEAK AI

- Weak AI is focused towards the technology which is capable of carrying out pre-planned moves based on some rules and applying these to achieve a certain goal.
- Applications of Weak AI **make humans feel** like the machines are acting intelligently (but they are not).

STRONG AI

- The principle behind Strong AI is that the machines could be made to think or in other words could represent human minds in the future.
- The applications of Strong AI will (someday) actually act and think just as a human.

Weak AI vs Strong AI

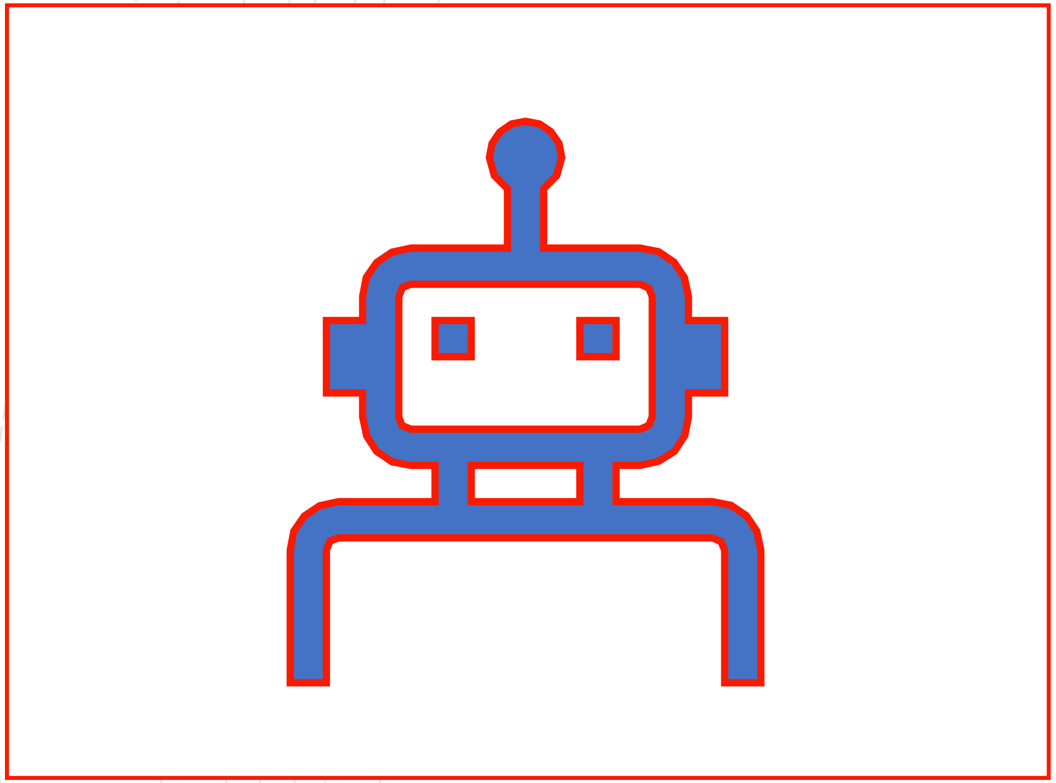
WEAK / NARROW AI

- When a human plays chess against a computer, the human may feel as if the computer is making impressive moves.
- However, all the moves that it makes are previously fed into the computer by a human and that is how it is ensured that the software will make the right moves at the right times.

STRONG AI

- Artificial General Intelligence
- In this case the AI would have the ability to actually think about the game and to plan chess moves in advance
- This form of AI would have the ability to control the game rather than simply react to the humans chess moves.

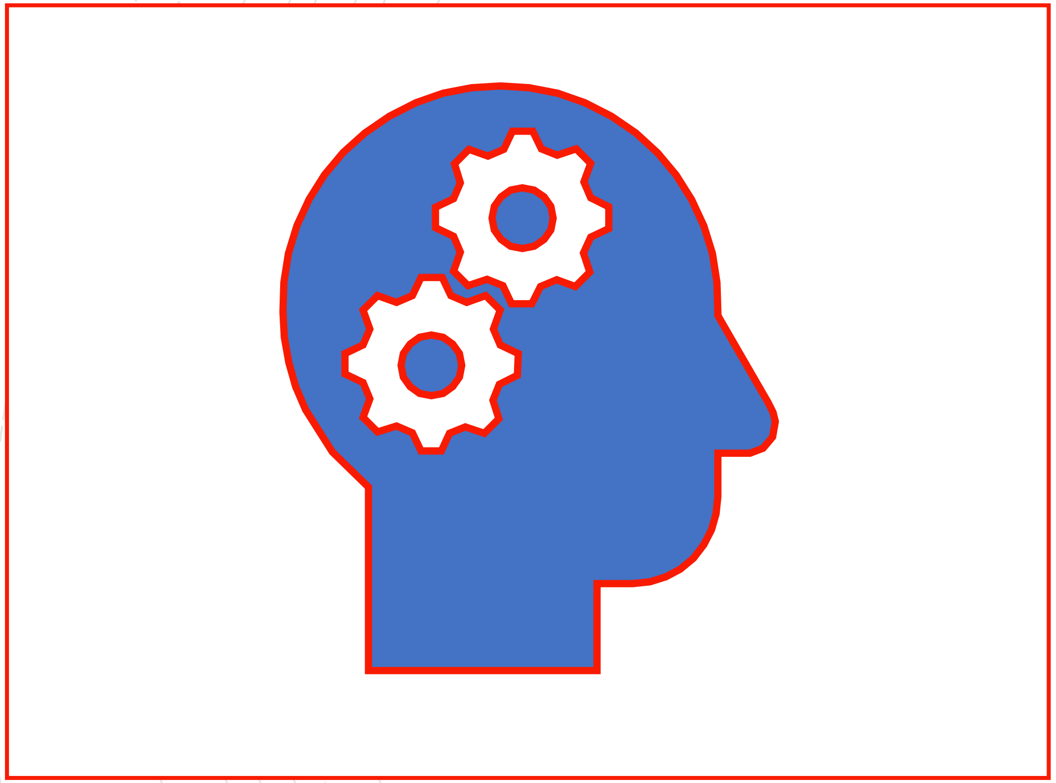
The challenges of Weak AI



- Weak AI or Narrow AI is focused on a specific task
  - *It is an advanced form of automation*
  - *It is usually paired with some optimization (find the best answer) or classification (what is it) task*

- Weak AI tends to be
  - *Easier to create*
  - *Stops working outside of expected range*
  - *Vulnerable to be misled or "stupid decisions"*

The challenges of Strong AI



- It is questionable if "strong/general AI" is possible?? Progress has been slow

- **We do not understand how it works in humans**
  - We cannot model an AI after our own brains as we do not understand our own brains (psychologically or biologically), or how any brain actually works.

- **Are we taking the wrong approach?**
  - The practice of abstraction, which people tend to redefine when working with a particular context in research, provides researchers with a concentration on just a few concepts. This is excellent for task-solving, but not for generalized intelligence.

Weak AI
vs
Strong AI

WEAK AI

- In this case, the ethics of the machine comes from the programmer

STRONG AI

- In this case the machine must develop/apply its own moral compass

The top 7 ethical issues in artificial intelligence

1. Humanity
   - How do machines affect our behaviour and interaction?
2. Robot Rights
   - How do we define the humane treatment of AI?
3. Biased AI
   - How do we eliminate AI bias?
4. Artificial Stupidity
   - How can we guard against mistakes?
5. Security
   - How do we keep AI from being misused?
6. Evil Genies
   - How do we protect against unintended consequences?
7. The Singularity
   - How do we stay in control of a complex intelligent system?

https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

1. Humanity

How do machines affect our behaviour and interaction?

Culture Shock

- Culture shock is a feeling of uncertainty, confusion, or anxiety that results from being cut off from your familiar environment and norms.
- When the way we live changes significantly this can have a big impact on people who do not adapt easily e.g. elderly people
- The idea of having an automated lifestyle would be quite alien to my father.
- This will not be a problem for your generation but by then there could be something new.

Culture Shock

- For example if the service industry were to be automated, the lives of the elderly in Ireland would change a lot.
- In Ireland, if elderly people are feeling lonely they will go to cafes, pubs, restaurants etc. Sometimes they will go to talk to the staff not just to meet friends.
- Having these places automated could make them feel lonely or stop them from attending because they don’t understand the technology.

What could we lose?

- Research on child-minding robots in the United States have demonstrated close bonding and attachment by children, who, in most cases, prefer a robot to a teddy bear.
- Short-term exposure can provide an enjoyable and entertaining experience that creates interest and curiosity.
    - Because of the physical safety that robot minders provide, children could be left without human contact for longer periods of time. The possible psychological impact of the varying degrees of social isolation on development is unknown.
- What would happen if a parent were to leave a child in the safe hands of a future robot caregiver almost exclusively?
    - The truth is that we do not know. We cannot conduct controlled experiments on children to find out the consequences of long-term bonding with a robot.
- We can get some indication from early psychological work on maternal deprivation and attachment.
    - Studies of early development in monkeys have shown that **severe social dysfunction occurs in infant animals** allowed to develop attachments only to inanimate surrogates

ROBEAR

- Japans prototype solution to an ageing population
- ROBEAR can lift and support patients
- This robot can also provide companionship to the elderly
- He can interact with his patients, encourage them to be active, remind them to take medication
- He can also allow medical professionals to log in and interact with the patients themselves.

People can get quite attached to their robots

- Robo Vacuums are quite popular
- Some people deliberately buy them with or give them faces
- In some cases, people treat them like pets instead of appliances and even give them names.

2. Robot Rights

How do we define the humane treatment of AI?

What is Moral Status?

- Francis Kamm has proposed the following definition of moral status, which will serve for our purposes: (Kamm 2007: chapter 7; paraphrase)

- X has moral status = because X counts morally in its own right, it is permissible/impermissible to do things to it for its own sake.

Moral status

- A rock has no moral status: we may crush it, pulverize it, or subject it to any treatment we like without any concern for the rock itself

- A human person, on the other hand, must be treated differently

- It is expected that we treat inanimate objects differently to animate objects.

Moral status

- Exactly what it means to treat a person correctly is something about which different ethical theories disagree; but it certainly involves
    - Taking their legitimate interests into account
    - Giving weight to their well-being
    - It may also involve accepting strict moral constraints in our dealings with them, such as a prohibition against murdering them, stealing from them etc.
- It is because a human person counts in their own right, that it is impermissible to do these things to them. This can be expressed more concisely by saying that a **human person has moral status**.

Moral status

- Moral behaviour is compliance with a set of ethics – of what is right/wrong
- **Can a machine be moral?**
- Our dealings with beings possessed of **moral status** are not exclusively a matter of rational functionality: we also have moral reasons to treat them in certain ways, and **to refrain from treating them** in certain other ways.

Moral status

- **Do we care about the machine as something other than as a machine**

- Humans love **anthropomorphism (or personification) -** the attribution of human traits, emotions, and intentions to non-human entities. It is considered to be an innate tendency of human psychology.

- *If you kicked a machine and broke it, does it feel like you wasted a machine or have you hurt someone?*

Intelligent Machines / AI

- **It is widely agreed that current AI systems have no moral status.**
- We may change, copy, terminate, delete, or use computer programs as we please; at least as far as the programs themselves are concerned.
- The moral constraints to which we are subject in our dealings with contemporary AI systems are all grounded in our responsibilities to other beings, such as our fellow humans, not in any duties to the systems themselves.
- *Do you feel uncomfortable about deleting large sections of a programme that you have written?*
- *Would you feel uncomfortable if it was the AI programme of a household robot that the family had bonded with?*

What would grant moral status??

- Two criteria are commonly proposed as being importantly linked to moral status, either separately or in combination:
- **Sentience**: is the capacity to feel, perceive, or experience subjectively, such as the capacity to feel pain and suffer
- **Sapience**: is the ability to think and act using knowledge, experience, understanding, common sense, self-reflection, and insight

What would grant moral status??

- Neither of these are possible today. Sapience is arguably closer than sentience. I think a sapient, non-sentient machine could be scary.
- So by definition, a machine today cannot be responsible for its decisions
- The phrasing of this might affect how people treat psychopaths…

So today…

- A machine is a reflection of the ethics and morality of the programmer and developer
- You are hard-coding a set of ethical rules into the system, therefore, they won’t evolve or adapt
- The programmers ethics, their sense of right/wrong is how the machine will behave
- No-one has precisely the same sense of values so it is not easy to agree, or even define the question as to what is the right thing to do
- Refusing to make a decision is still a decision

Something to think about…

- If two beings have the same functionality and the same consciousness experience, and **differ only in how they came into existence**, then they have the same moral status.

- So logically…

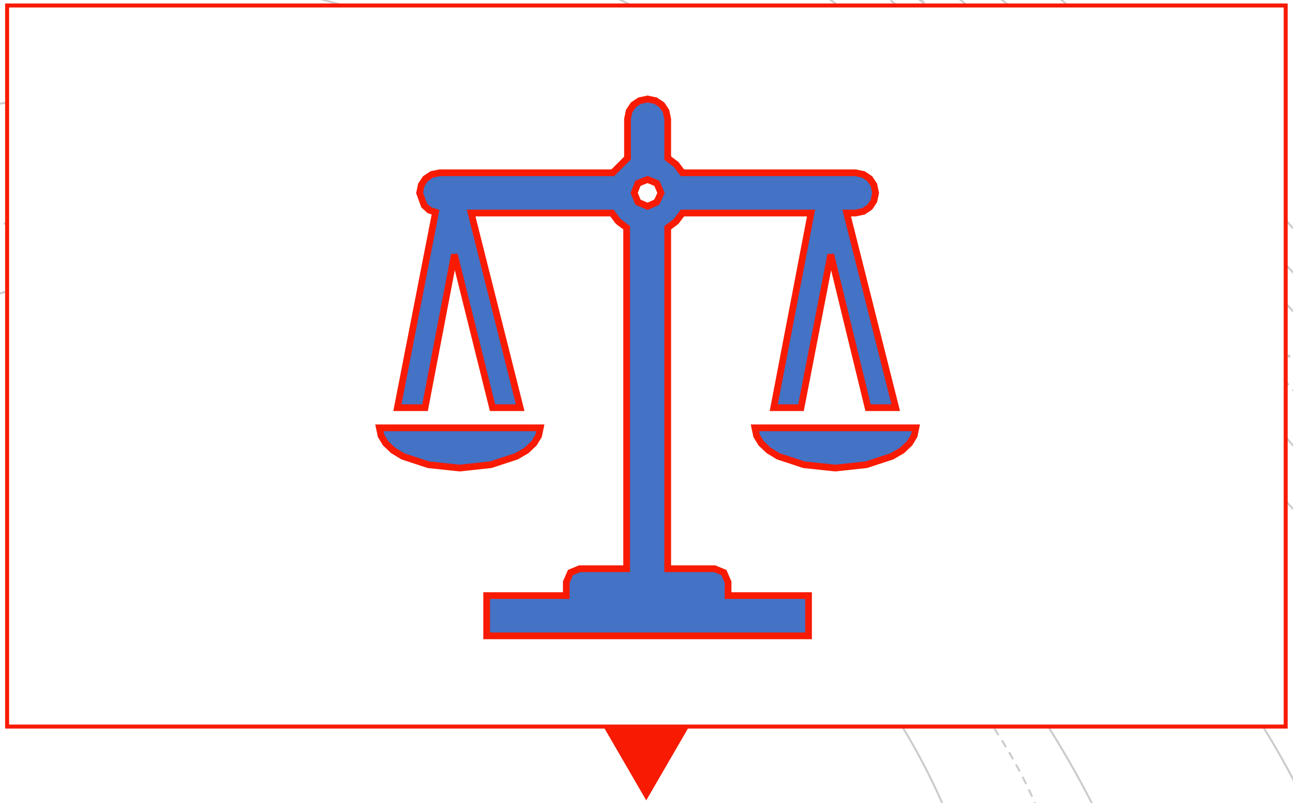
- If a machine and a human are functionally the same (from a behavioural perspective), then they should be given the same moral status as each other – irrespective of one being born and the other made.

- **If not legally, then morally, the machine and the human are the same.**

- **Remember Sophia from section 4?**

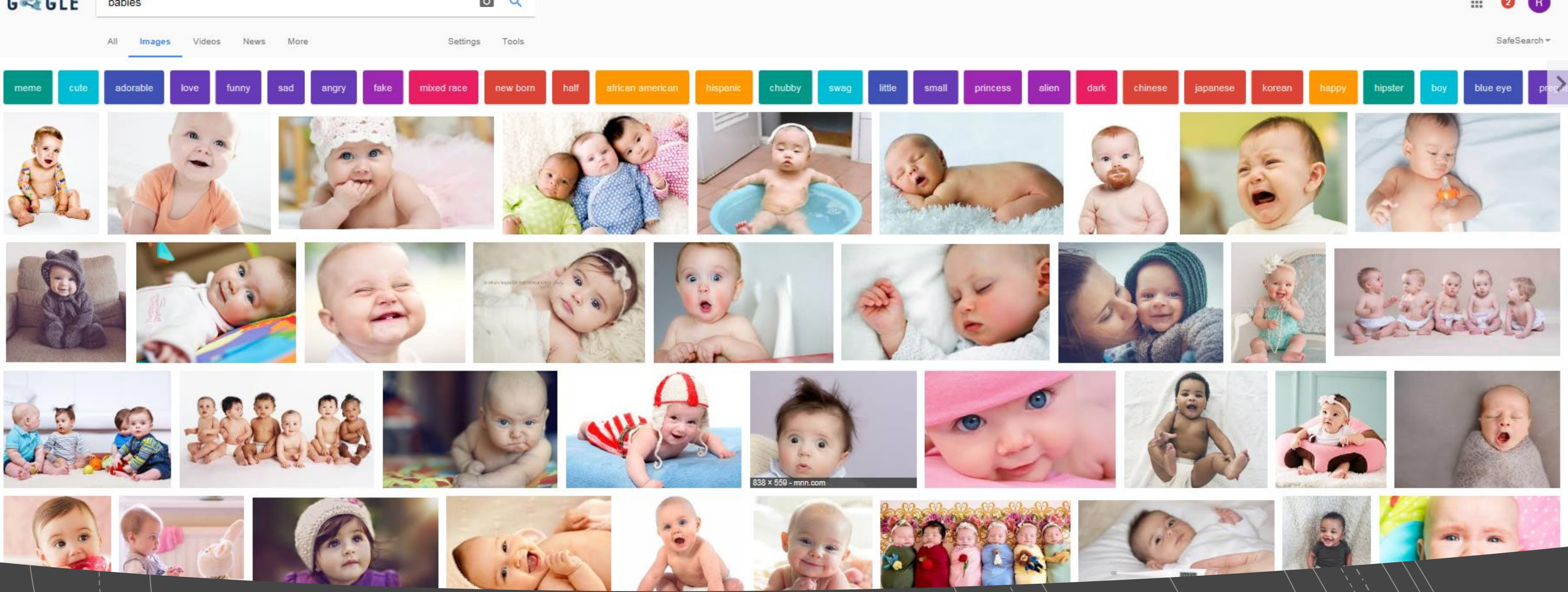
3. Biased AI

How do we eliminate bias in AI systems?



Bias and Ethical Framework

Bias is your conscious or sub-conscious interpretation of data to reflect your world view

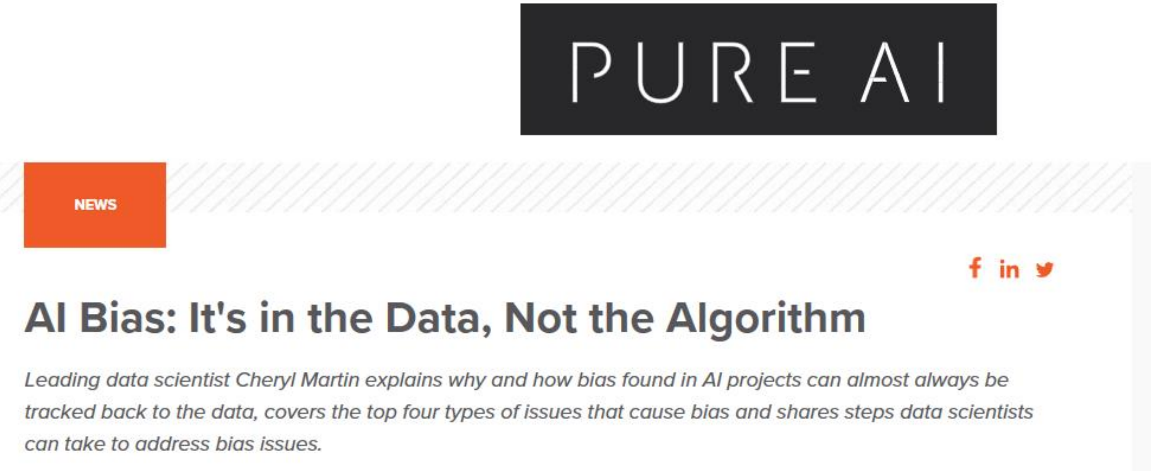


Racist robots: How do we eliminate bias

- **GOOGLE: Only 3 babies out of 69 are not-white? Why?**
- **Roughly 75% of babies should be non-white**

Racist robots: How do we eliminate bias

- Are we un-knowingly programming in prejudices?
- Is our training data flawed?
- Are we "indoctrinating" our creations with our biases through our algorithms or training experiences? (Nature and Nurture)
- We can see the obvious prejudices, can we see the subtle ones?

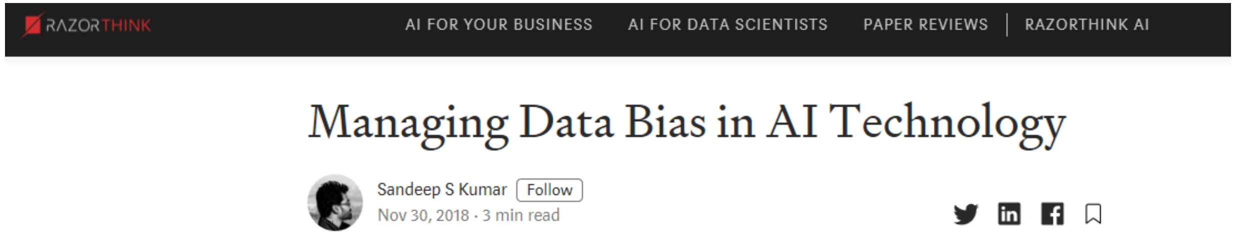


PURE AI

NEWS

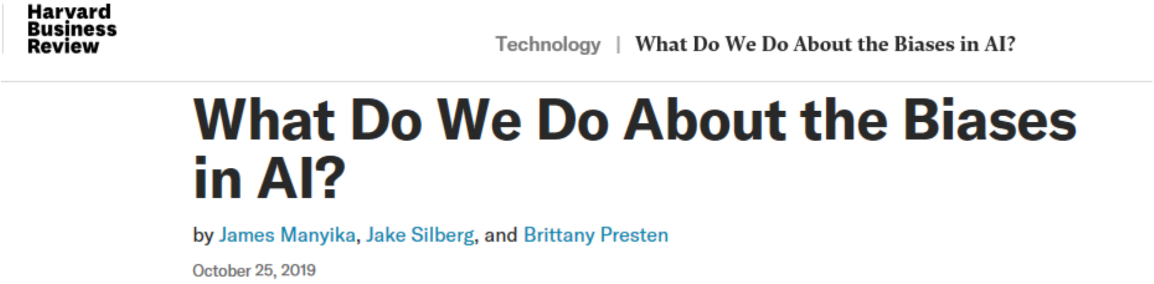
AI Bias: It's in the Data, Not the Algorithm

Leading data scientist Cheryl Martin explains why and how bias found in AI projects can almost always be tracked back to the data, covers the top four types of issues that cause bias and shares steps data scientists can take to address bias issues.



Managing Data Bias in AI Technology

Sandeep S Kumar Follow
Nov 30, 2018 · 3 min read



What Do We Do About the Biases in AI?

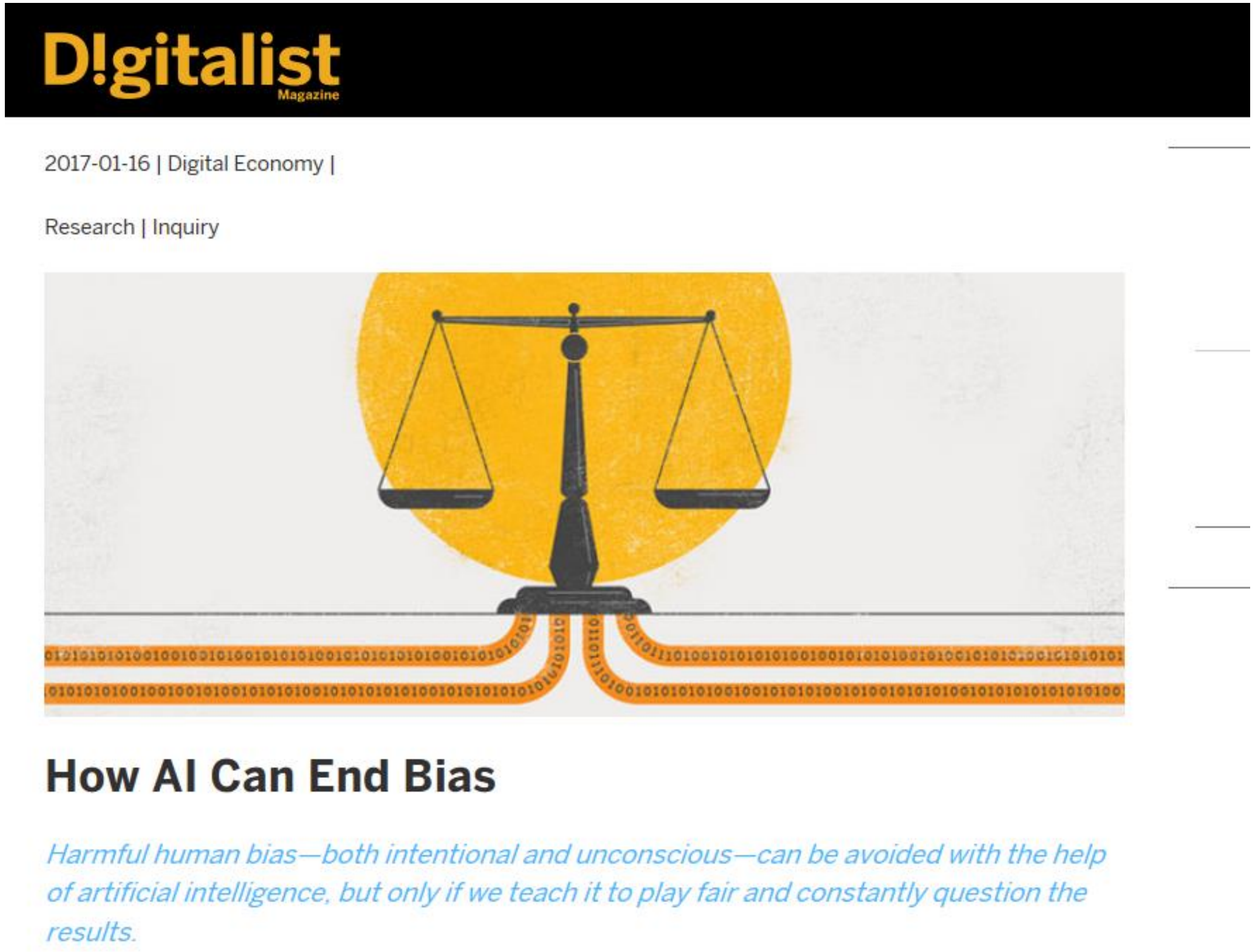
by James Manyika, Jake Silberg, and Brittany Presten

October 25, 2019

AI Bias in the News?

- *Technologist Kriti Sharma recently pointed out that the first wave of virtual assistants reinforced sexist gender roles: the "**personal assistant**" roles such as Apple's Siri and Amazon's Alexa have **female** voices, but "**problem-solving**" bots like IBM's Watson and Microsoft's Einstein have **male** ones.*

AI Bias in the News?



D!gitalist Magazine

2017-01-16 | Digital Economy |

Research | Inquiry

How AI Can End Bias

Harmful human bias—both intentional and unconscious—can be avoided with the help of artificial intelligence, but only if we teach it to play fair and constantly question the results.

COMPAS

WHAT IS IT?

- Correctional Offender Management Profiling for Alternative Sanctions
- Essentially it a tool that was designed to allow law enforcement to autonomously evaluate criminals by a ranking system

POSSIBLE EVIDENCE OF BIAS

- COMPAS has been the center of controversy on a number of occasions
- The system has been accused of considering white people less of a criminal risk than other ethnicities
- In cases where similar crimes have been committed the system appears more likely to assign a white offender a lower rank.

COMPAS

Brisha Borden

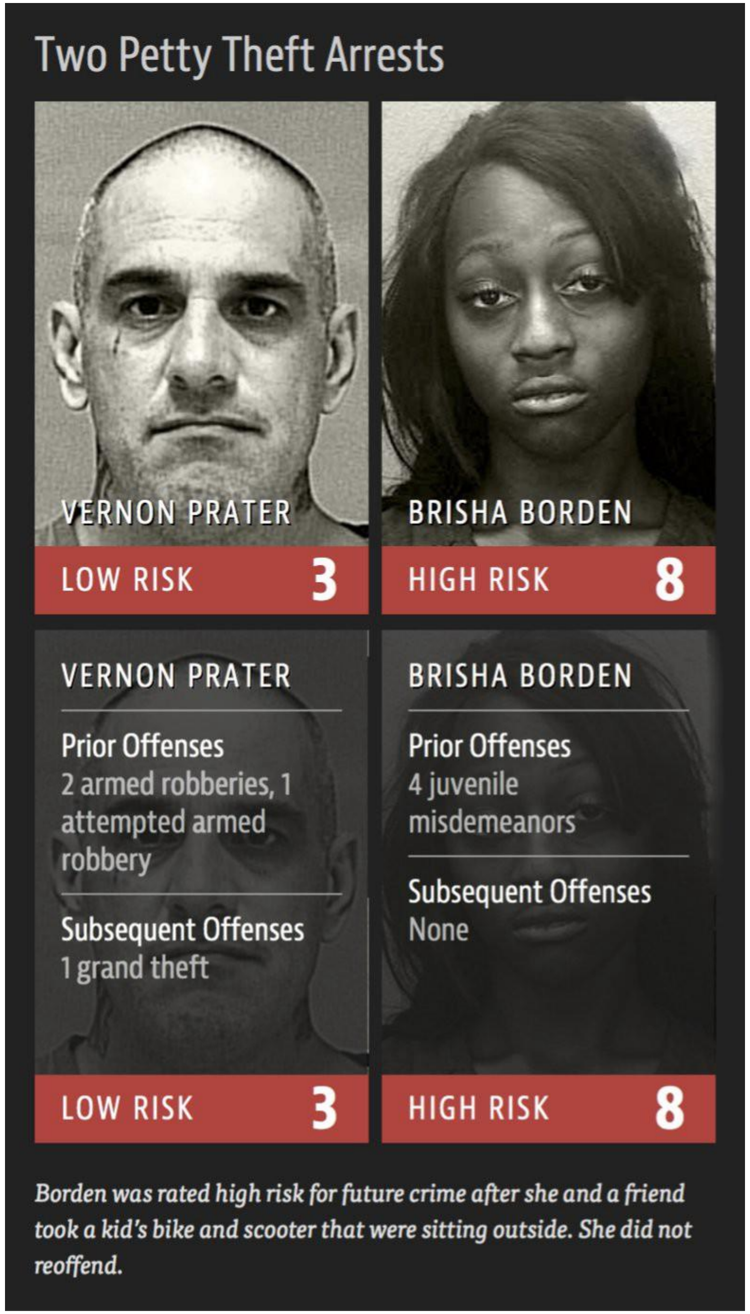
- Brisha Borden was 18 and had a number of misdemeanors on record from when she was a juvenile
- She stole a bicycle that was lying on the street worth approximately $80
- She did drop the bicycle when the owner saw her and shouted at her to stop
- However, she was arrested and charged with theft
- COMPAS ranked Brisha as 8 i.e. that she was a high risk offender and was very likely to reoffend
- However, 2 years later Brisha had not committed any other crimes

COMPAS
Vernon Prater

- Vernon Prater was 41 he had already been convicted twice for armed robbery and had previously spent 5 years in prison
- He was arrested for stealing 90$ worth of tools from a hardware store
- COMPAS ranked Prater as a 3 i.e. a low risk offender who would be unlikely to commit another crime
- However, 2 years later he had been arrested again for armed robbery and was serving an 8 year prison term

Why were these ranks chosen?

- One was proven to be a violent re-offender and the other wasn't.
- Does the system favour white people or was there some other reason?
- Is it based off Americas 3 strike system?
- Could the ranking be based of offender statistics from where they live?
- This can be the problem with 'black box' algorithms
- There are a number of examples of this e.g. Wisconsin vs. Loomis where he said using COMPAS violated due process because the process was not transparent



Two Petty Theft Arrests

VERNON PRATER	BRISHA BORDEN
LOW RISK 3	HIGH RISK 8

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

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.

Credit Pro Publica

Explainable/Transparent AI

- Explainable AI (XAI) is a growing trend… it’s a reaction to opaque/black-box decision making algorithms – where not even the creators know why the machine works in the way it does – just that it typically does. (neural networks are particularly bad about this)

- The objective is that the results of the solution can be understood by human experts.

- Some argue that Explainable AI principles may become a legal requirement as not knowing how a decision was made is just too opaque and resistant to inspection/correction.

- (this is particularly relevant around social or ethical systems)

IEEE Global Initiative

- At the end of these notes we will look more closely at a global initiative from the IEEE
- Global initiatives from technical institutions like the IEEE could help significantly reduce bias in AI
- Having multiple ethnicities, ethical frameworks, and resources contributing to creating AI is a good thing
- It reduces the risk of a single person/source/viewpoint being able create AI that serves only their own purpose and not the good of everyone
- Some of the solutions that have been looked at are on the following slides

Solutions

1

Machine - algorithms will have different biases than humans. They can be used to monitor each other to detect biases.

2

Engage in fact-based conversations around potential human biases. When we do find bias, it is not enough to change an algorithm—you need to change the human processes that caused the bias.

3

Have transparency about the algorithms so you can explain how a decision was made.

4

Diversify the data sets and the algorithms. Different groups / data-sources will have different biases.

Multiple inputs should help highlight biases or inconsistencies.

4. Artificial Stupidity

How can we guard against mistakes?

Turing Test

WHAT IT IS

- The Turing test was developed by Alan Turing in 1952 as a test for Artificial Intelligence
- The idea was that someone would ask two ‘people’ questions and try to determine if they were a real person or a machine

HOW IT CAN BE BEATEN

- The easiest way to find the machine was to ask it math questions. The machine would give very accurate answers that a human couldn’t
- In this way, it was determined that the computer should make deliberate mistakes to appear more human e.g. a chatbot making spelling mistakes

Fabio the Pepper Robot

THE PEPPER ROBOT

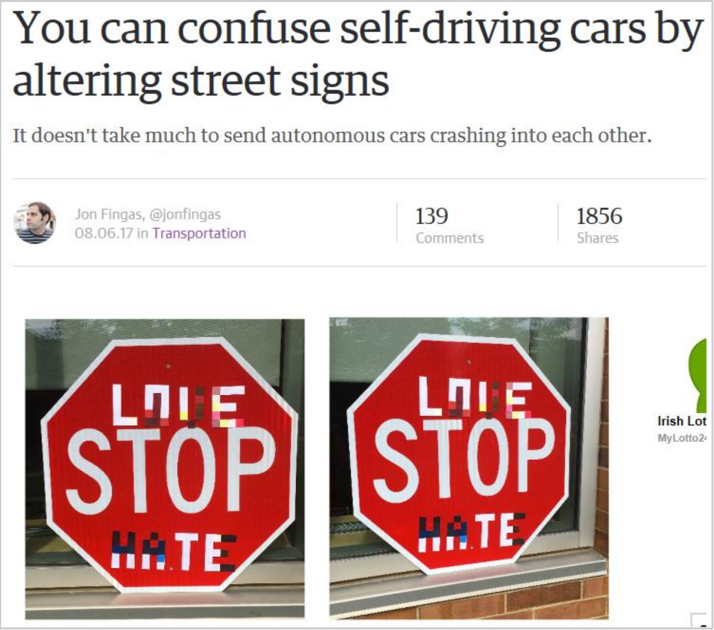
- The pepper robot was originally developed as a companion type robot. Designed to help their users have fun and enjoy life
- They were used to great success in many places as companions for the elderly, receptionists, and store greeters

FABIO

- Fabio was used in a Scottish supermarket as a greeter and to give customers information about the store
- Fabio would shout 'hey, gorgeous' at customers and initiate unwanted physical contact
- They had to remove him because customers started avoiding him and the shop

Can you identify this object?

Deliberate ‘mistakes’



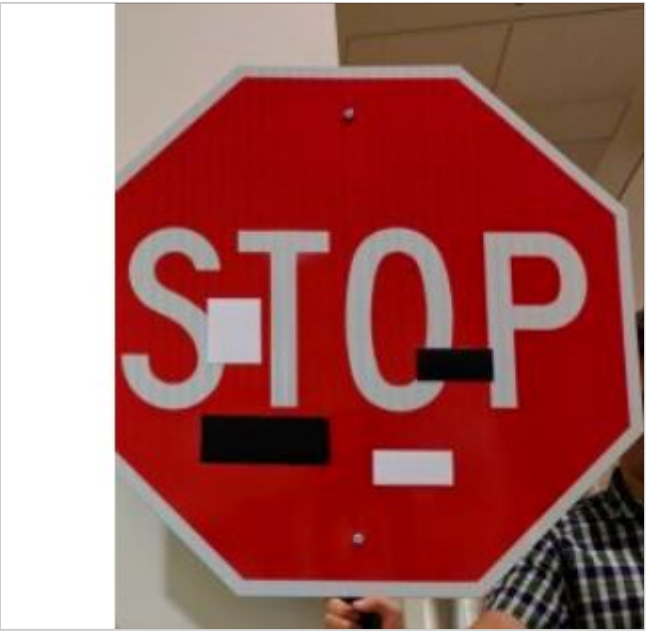
You can confuse self-driving cars by altering street signs

It doesn't take much to send autonomous cars crashing into each other.

Jon Fingas, @jonfingas
08.06.17 in Transportation

139 Comments

1856 Shares



- Machines are trained against a set of experiences. This cannot be exhaustive.
- Machine Systems can be fooled in ways that humans wouldn't be.
- For example, random dot patterns can lead a machine to “see” things that aren’t there.
- We need to ensure that the machine performs as planned, and that people can’t confuse it – accidentally or maliciously

5. Security

How do we keep AI from being misused?

Remember Maroochy Beach?

- Case study in Notes Section 3
- An angry employee misused the system to dump large amounts of waste into the beach and surrounding areas
- It is important to not only have failsafes but to also design a robust system

Baymax from Big Hero 6

- Baymax was originally a healthcare robot
- His owner, Hiro, altered him to be a combat robot that added to his original programming
- When the original programming was tampered with, the robot injured a number of allies and was difficult to stop.
- What if this was done intentionally by someone who wanted to perform criminal acts?

Credit to PIXAR and Big Hero 6

Solutions

WHITE HAT HACKING

- White hat hackers will deliberately try to break, misuse, or repurpose software in order to report it
- By using these types of hackers to search for faults in a system it prevents black hat hackers for exploiting these weaknesses

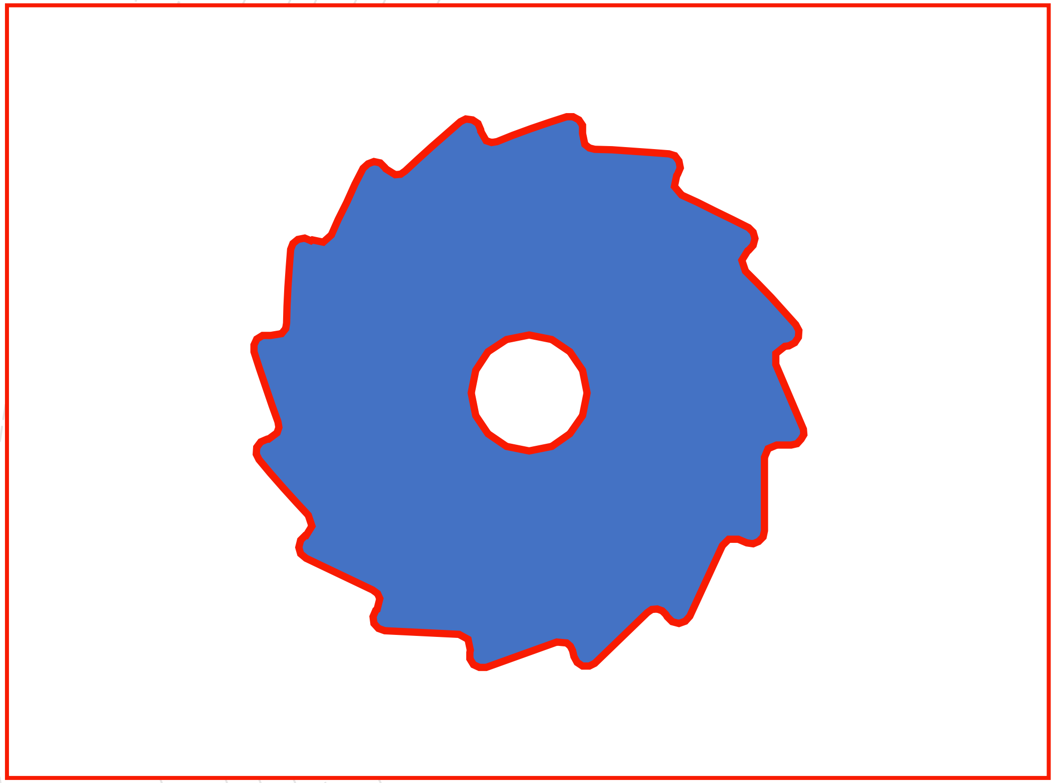
ROBUSTNESS TESTS

- A type of idiot test where people will deliberately try to break, misuse, or repurpose machines.
- This simulates people who might cause a machine to malfunction in a dangerous way though malice or stupidity and then exploit this for their own purpose

6. Evil Genies

How do we guard against unintended consequences?

If the machines that create our machines break down, will we still have the skills to fix them?



- *"Let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind.* ***Thus the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.*** *It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously."*

- Irving John Good, 1960's

How do we protect against unintended consequences?

- We can imagine an advanced AI system as a "genie in a bottle" that can fulfill wishes, but with terrible unforeseen consequences.
- There is unlikely to be malice at play, only **a lack of understanding of the full context** in which the wish was made.
- Remember the hiring algorithm that preferred to hire males? Or the algorithm that predicted that people over 100 had lower mortality rates from pneumonia?
- There are many cases were "optimisation" may lead to subtle but undesirable outcomes

An Extreme Example

Kurt Vonnegut
Cat’s Cradle

- A fictional novel about a scientist that invents a way to turn water into a solid at room temperature
- The way it does this is through a molecule called ice-nine that will cause any body of water that it is introduced into to freeze
- The invention could have been used to transport water to areas of drought or to stabilize the ice shelves of the arctic caused by global warming
- After the scientist died, his children divided up samples of the molecule to profit from them
- Fighting between the children ultimately led to the plane transporting one of the last samples being shot down over the sea, causing the worlds water supply to freeze.

It is difficult to put a genie back in a bottle…

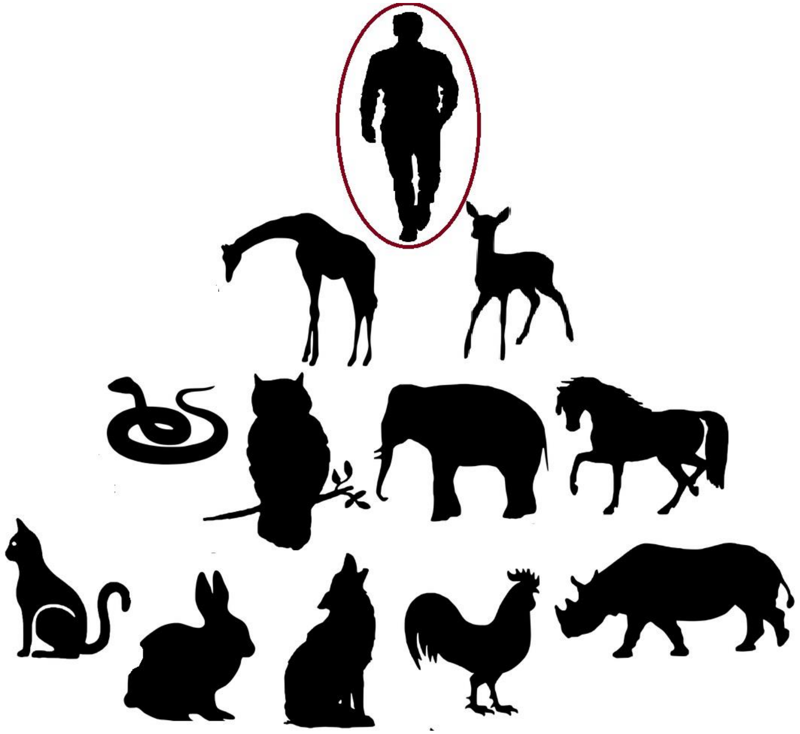
- The children in the story did not fully understand or appreciate the consequences of releasing the molecule
- This returns to the idea of unintentional design error and the steps that must be taken to prevent it
- However, in this case advertising ‘do not order your robot to steal’ on the packaging is not sufficient
- An autonomous machine that is free to do as it pleases or that has the ability to do something potentially dangerous by itself could be a big problem

7. The Singularity

How do we stay in control of a complex intelligent system?

Top of the Food Chain?

- The reason humans are on top of the food chain is not down to sharp teeth or strong muscles. Human dominance is almost entirely due to our ingenuity and intelligence.
- This poses a serious question about artificial intelligence: will it, one day, have the same advantage over us? We can't rely on just "pulling the plug" either, because a sufficiently advanced machine may anticipate this move and defend itself.
- This is what some call the “singularity”: the point in time when human beings are no longer the most intelligent beings on earth.



Luciano Floridi, Philosopher of Information

- ***How some nasty ultra-intelligent AI will ever evolve autonomously from the computational skills required to park in a tight spot remains unclear***

Friendly AI

- If research into strong AI produced sufficiently intelligent software, it would be able to reprogram and improve itself – a feature called "recursive self-improvement".
- It would then be even better at improving itself, and would probably continue doing so in a rapidly increasing cycle, leading to an intelligence explosion and the emergence of superintelligence.
- Such an intelligence would not have the limitations of human intellect, and might be able to invent or discover almost anything.

- Hyper-intelligent software **might not necessarily decide to support the continued existence of mankind, and might be extremely difficult to stop**. This topic has also recently begun to be discussed in academic publications as a real source of risks to civilization, humans, and planet Earth.

Wikipedia

Friendly AI

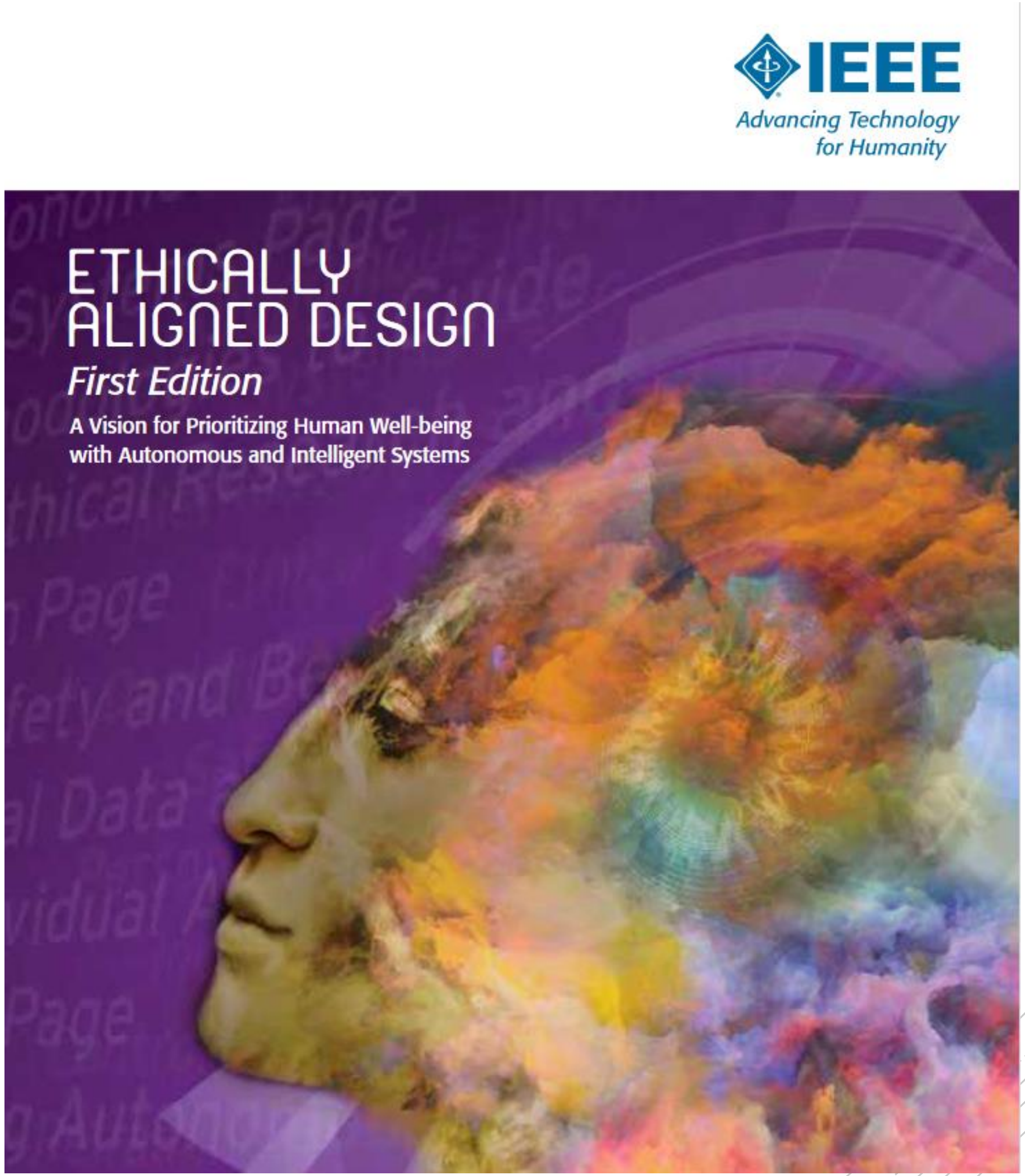
- One proposal to deal with this is to make sure that the first generally intelligent AI is a friendly AI that would then endeavor to ensure that subsequently developed AIs were also nice to us. But friendly AI is harder to create than plain AGI, and therefore it is likely, in a race between the two, that non-friendly AI would be developed first. Also, there is no guarantee that friendly AI would remain friendly, or that its progeny would also all be good.[89]

- Yudkowsky (2008) goes into more detail about how to design a **Friendly AI**. He **asserts that friendliness (a desire not to harm humans) should be designed in from the star**t, but that the designers should recognize both that their own designs may be flawed, and that the robot will learn and evolve over time.

- Thus the challenge is one of mechanism design—**to define a mechanism for evolving AI systems under a system of checks and balances,** and to give the systems utility functions that will remain friendly in the face of such changes.

Where are we now?

IEEE Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems



IEEE
Advancing Technology
for Humanity

ETHICALLY
ALIGNED DESIGN

First Edition

A Vision for Prioritizing Human Well-being
with Autonomous and Intelligent Systems

https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined

What is it?

- The goal of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (“The IEEE Global Initiative”) is that Ethically Aligned Design will provide pragmatic and directional insights and recommendations, serving as a key reference for the work of technologists, educators and policymakers in the coming years.

- Ethically Aligned Design sets forth scientific analysis and resources, high-level principles, and actionable recommendations. It offers specific guidance for standards, certification, regulation or legislation for design, manufacture, and use of A/IS that provably aligns with and improves holistic societal well-being.

General Principles as Imperatives

1. **Human Rights** – create and operate to respect, promote, and protect internationally recognized human rights.
2. **Well-being** – adopt increased human well-being as a primary success criterion for development.
3. **Data Agency** – empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
4. **Effectiveness** – provide evidence of the effectiveness and fitness for purpose of A/IS.
5. **Transparency** – the basis of a particular A/IS decision should always be discoverable.
6. **Accountability** – create and operate to provide an unambiguous rationale for all decisions made.
7. **Awareness of Misuse** – guard against all potential misuses and risks of A/IS in operation.
8. **Competence** – specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

A/IS creators an operators shall….

A/IS = Artificial or Intelligent Systems

Statement on Ethics for Intelligent Machines From the Royal Irish Academy

https://www.ria.ie/sites/default/files/35201_ria_booklet_robotics_and_personhood.pdf

RIA Policy Statement

- **Public policy and experience-centred design**
- What we allow robots to do and where we give primacy to human activity and experience is clearly a matter for policy. This has recently been indicated by the European Parliament Committee on Legal Affairs' (2016) call for a guiding ethical framework for the design, production and use of robots and the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (2019) the purpose of which is to '***ensure every technologist is educated, trained, and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems***'.

- An immediate focus for policy should be a clear articulation of the values of robotics design practice, an ethical approach to which would be informed and motivated by a concern for human experience and the ways in which people make sense of it.

RIA - Care

- Caring, which is a process of protecting and looking after the needs of another, and often requires intimate communication of needs, concerns and recognition.

- Care requires an **empathic relationship** in which people mutually accept each other and are responsive to each other's feelings. That does not mean that robots cannot be used to entertain, stimulate, educate, protect or foster connections between people. Rather, their use in these contexts **should be subject to an overriding concern for the primacy of human contact in care, the minimum requirement to sustain personhood**

RIA - Summary

- We are beginning to see a growing and welcome focus on the ethics of robotics and on the user-perceived quality of interactions with robots. Over time, and with political will, public policy and law may provide a protective framework for human–robot interaction. Over time, people will also learn about the potential and the limitations of social robots.

- In the meantime, if only to ensure that the potential for beneficial development of robotics is not hampered, **robotics design should give due weight to personhood, human activity and experience.**

Thought Experiments

Thought Experiment 1

- You are making an automated drug injection device. You have a number of sensors monitoring your patient

- Some of the sensors indicate that you need to inject the medicine
- Some of the sensors indicate that you should not.

- What do you do in the presence of imperfect information?

- **No decision is still a decision.**

Thought Experiment 2

- You are working on an industrial gas burner capable of heating up 2000 tons of sulphuric acid to near-boiling point. If it gets too hot, you will make an acid cloud capable of destroying plant life for several miles. If you shut it down unnecessarily, you will lose your company millions of euro. (real scenario)

- You have a faulty temperature sensor, but only one of many. **You should be okay**.

- Do you continue operating? Or do you shut down?

Thought Experiment 3

- You are making an educational toy that gives out sweets if the child completes the maths puzzles

- The device is encouraging educational development which is a good thing
- It is rewarding with sweets which is sugary, bad for teeth, and probably contributes to obesity – which is a bad thing

- **When does a good thing turn into a bad thing?**

- Note: any reward-based system suffers from this issue