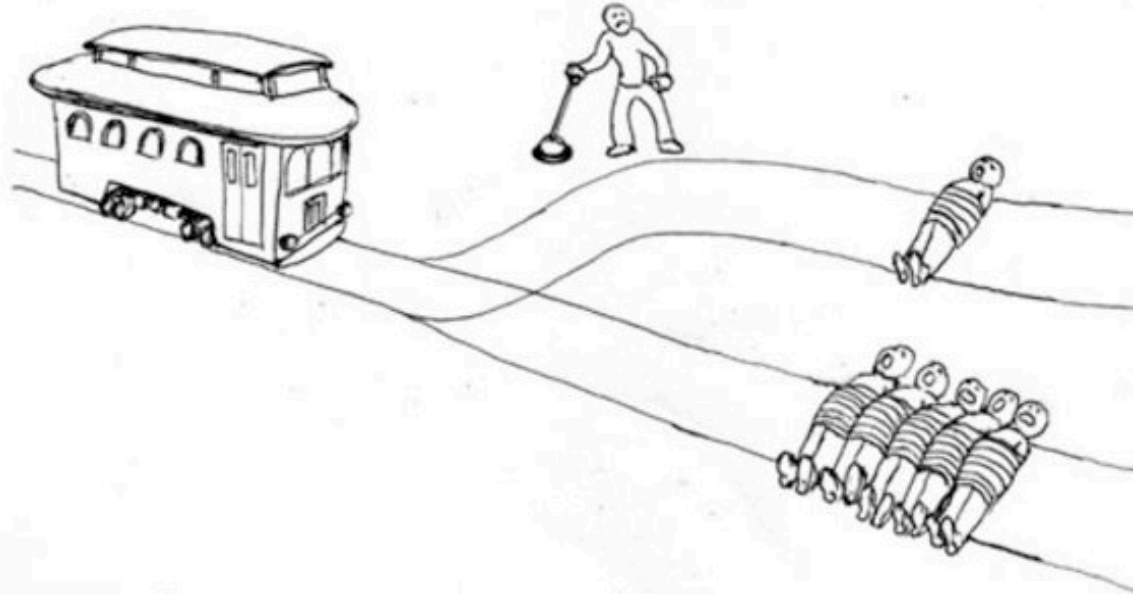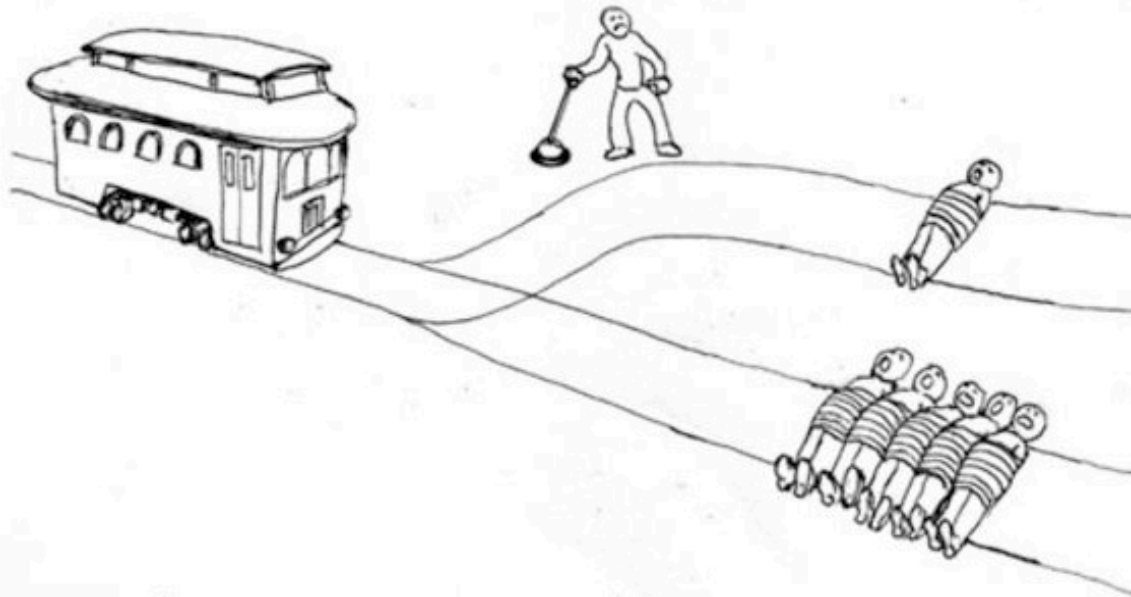what we *really* mean
when we say *ethics*
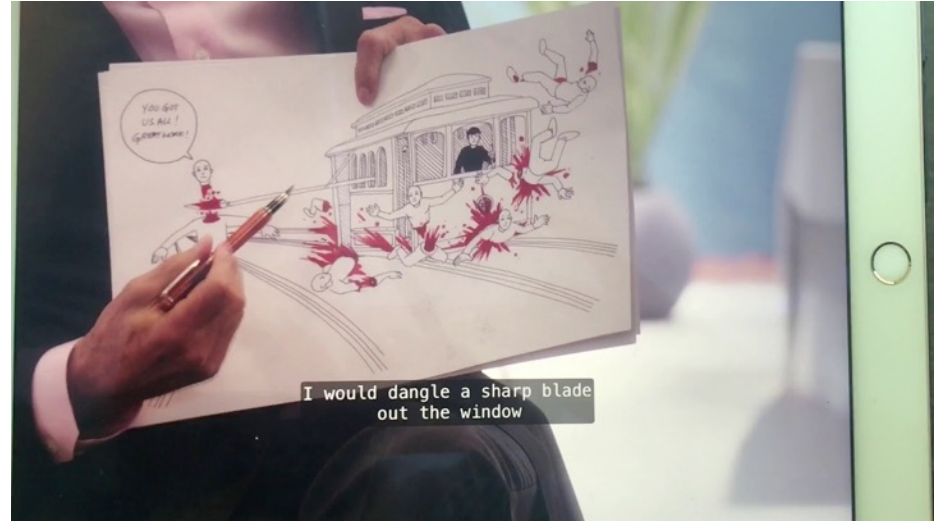
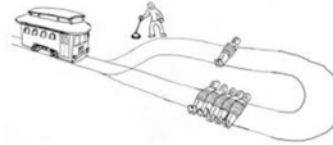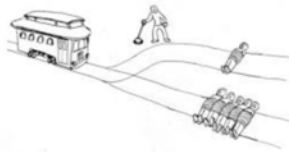girlwonder.com | @maximolly

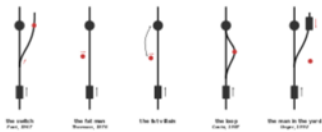Phillipa Foot (1967)

Judith Jarvis Thomson (1985)

PROBLEM SOLVED

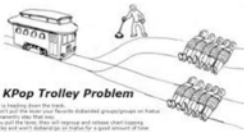Do you push the fat man to prevent a trolley problem?

The trolley problem of
Isaac Asimov's first law of robotics

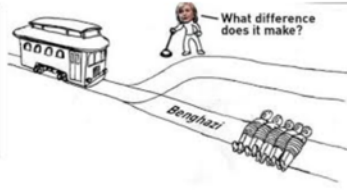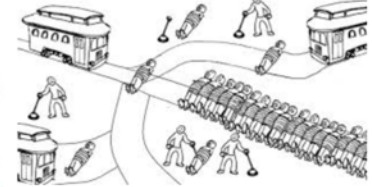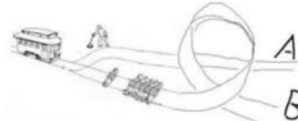1) A robot may not injure a human being or, through inaction, allow a human being to come to harm

injure a human being

allow a human being to come to harm

The KPop Trolley Problem

Hedonist's Trolley Problem

The track is heading towards B.

If you pull the lever, it will switch to A but it won't do the totally sick loop-da-loop.

A
B

What difference does it make?

Benghazi
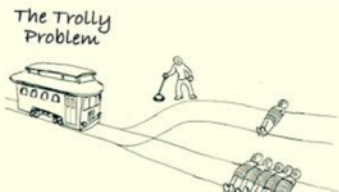
The Trolly Problem

The Existential Trolley Problem

No matter what you choose, you're gonna die in the end
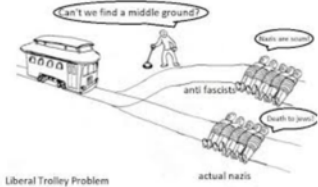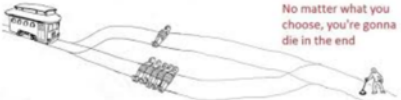
Can't we find a middle ground?

Nazis are scum

anti fascists

Death to jews!

Liberal Trolley Problem

actual nazis
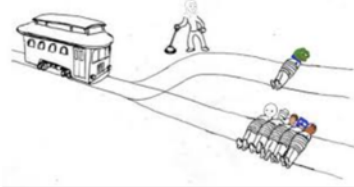
The Mutually Assured Trolley Problem
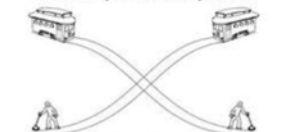
You and some other guy are glued to the tracks. Any of you can pull the lever, releasing the trolley and killing the other guy. You told the other guy that you'll pull the lever if he does, hoping he won't kill you. After a long time, he finally pulls the lever. Do you keep your promise and lead to 2 deaths instead of 1?

How you imagine the trolley problem
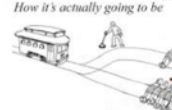
How it's actually going to be

Ship of Theseus Trolley Problem

If you do not pull the lever, one person will be crushed to death instantly. If you do pull the lever, the trolley will divert onto a thousand-mile stretch of track with one person tied down at the end of it. If as the trolley rolls down this thousand-mile track, a crew systematically switches out every piece of the original trolley with a replacement part, did the trolley which you diverted kill the man?

1000 miles
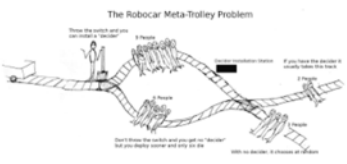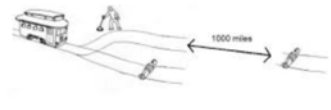
The Robocar Meta-Trolley Problem

Trolley problem memes

@Marcos Abenante

EVER HEARD OF THE TROLLEY PROBLEM?

NO. WHAT IS IT?

A TROLLEY IS BARRELING TOWARDS FIVE HELPLESS PEOPLE ON THE TRACKS. YOU CAN PULL A LEVER TO DIVERT IT ONTO ANOTHER TRACK, BUT—

CAN I REACH THE LEVER WITHOUT GETTING UP? LOOK, I'M NOT—

IN THIS SCENARIO HOW BUSY AM I?

I GUESS I FORGOT WHO I WAS TALKING TO.

FOR A DOLLAR, I'LL PROMISE TO PULL THE LEVER IF ONE OF THE FIVE PEOPLE IS YOU.

Trolley salesman: *slaps roof of trolley*
- This bad boy can fit so much ethics in it
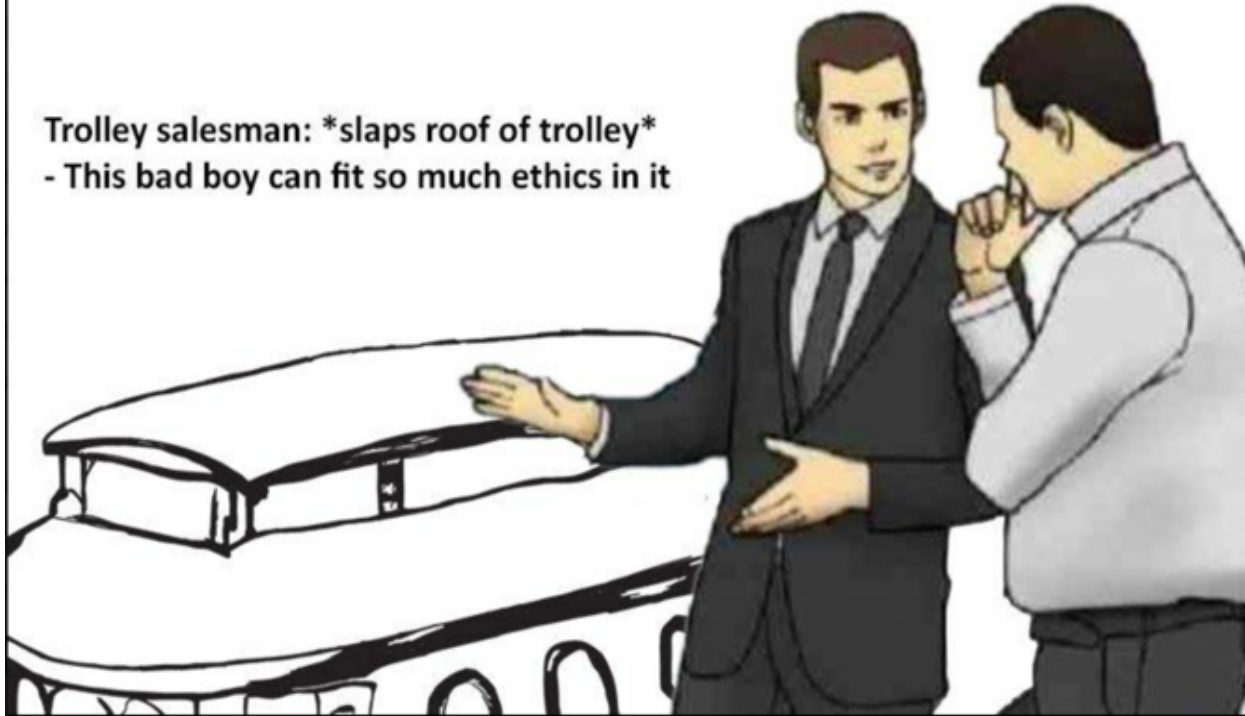
Markkula Center for Applied Ethics at Santa Clara University

Markkula Center for Applied Ethics at Santa Clara University

WHAT ETHICS IS NOT

FEELINGS

~~FEELINGS~~

RELIGION

~~RELIGION~~

LAW

LAW

SCIENCE

SCIENCE

SCIENCE

Markkula Center
for Applied Ethics
at Santa Clara University

WHAT ETHICS IS NOT
WHAT ETHICS IS

5 approaches:
Utilitarian
Rights
Fairness/Justice
Common Good
Virtue

5 approaches:
Utilitarian

5 approaches:

Rights
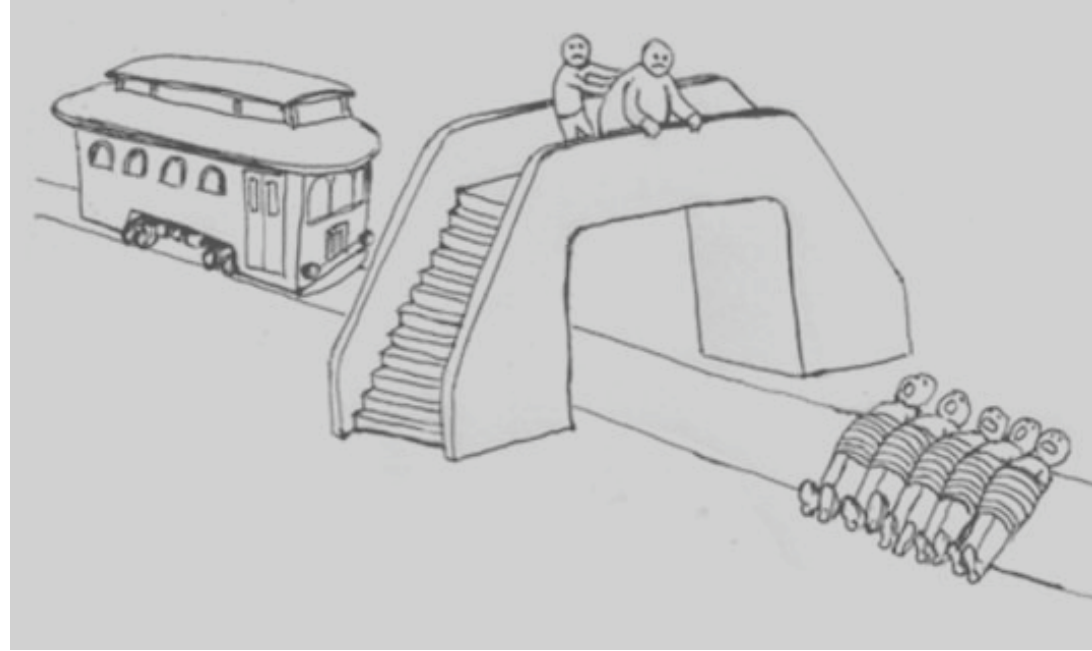
5 approaches:

Fairness/Justice

5 approaches:

"Sweetheart, could you maybe include the dog?"

Common Good

5 approaches:

Virtue Ethics Trolley Problem

As you're experiencing the trolley problem you ask yourself: "What would that chill guy I met at the bar do?"
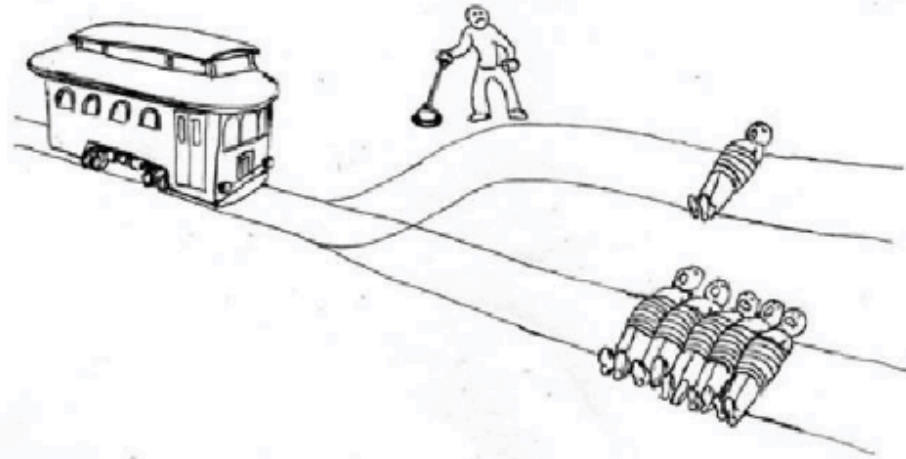
Virtue

2018: the year of ethics

Since 1/1/2018: more than 9,430 news articles on AI + ethics (Lexis Nexis)

2017: 3,878
2016: 2,064
2015: 1,160
2014: 725
2013: 451
2012: 360
2011: 311

so why **now?**

At least 50* different ethics...

Frameworks

Toolkits

Principles

Codes of conduct

Checklists

Oaths

Manifestos

(via Kathy Baxter, Architect of Ethical AI Practice at Salesforce, who said in a personal communication that the number may be closer to 200)

ETHICAL OS

a guide to
**anticipating
the future impact of
today's technology**

Or: how not to regret
the things you will build

Paula Goldman Joins Salesforce as VP,
Chief Ethical and Humane Use Officer

Framework for Trustworthy AI

CHAPTER I

Ethical Purpose

Ensure respect of fundamental rights, principles and values when developing, deploying
and using AI

Realisation of Trustworthy AI

Ensure implementation of ethical purpose as well as technical robustness when
developing, deploying and using AI

CHAPTER II

Requirements for Trustworthy AI

To be continuously evaluated, addressed and assessed in the Design & Use phase of AI
through

Technical Methods   **Non-Technical Methods**

CHAPTER III

Assessment List for Trustworthy AI based on Use Cases

ASILOMAR AI PRINCIPLES

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

Of oaths and checklists

Oaths have their value, but checklists will help put principles into practice.
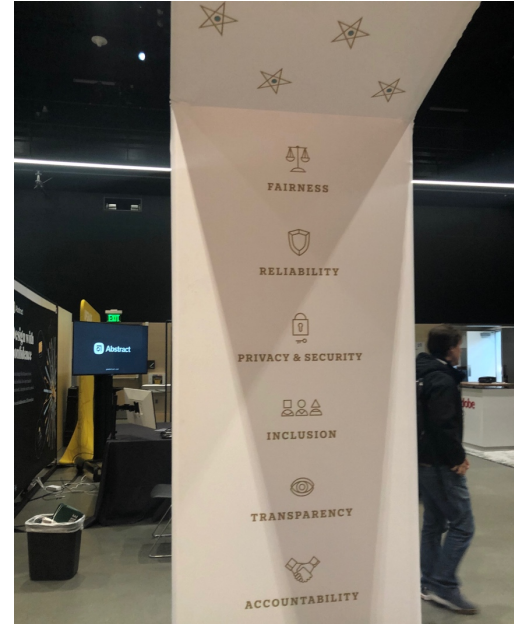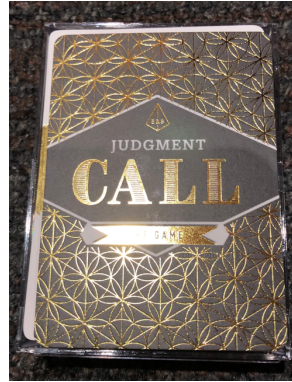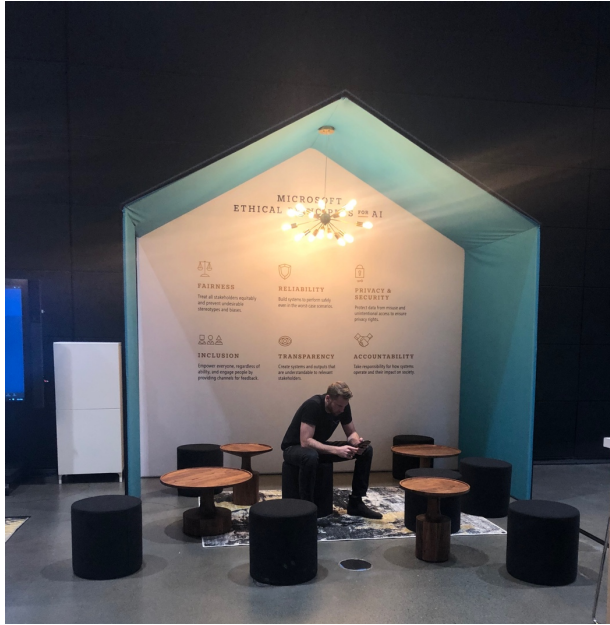
josh lefevre & louise larson
master's students
carnegie mellon school of design

kathy baxter
architect of ethical AI practice
salesforce

ethical frameworks are a **good thing**

but how do we know if they're **any good?**

or if they're even **ethics?**

many ethics frameworks don't include a **definition** of ethics

these tools **build community**

these tools **provoke thought**

these tools offer guidelines

fairness

privacy

safety

protection against unintended bias

protection against discrimination

"As a designer, I feel like my job is **covering technology's ass.**"

—Ruth Kikin-Gil

design is where the **rubber** meets the **road**

how can design **help**?

1. design is good at **framing problems**

2. designers investigate the **context** of a problem with human-centric approaches

3. what and how **data** is to be collected is a design question

Mimi Onuoha / The Library of Missing Datasets

~~Civilians killed in encounters with police or law enforcement agencies~~ [update: this is no longer a missing dataset]

Sales and prices in the art world (and relationships between artists and gallerists)

People excluded from public housing because of criminal records

Trans people killed or injured in instances of hate crime (note: existing records are notably unreliable or incomplete)

Poverty and employment statistics that include people who are behind bars

Muslim mosques/communities surveilled by the FBI/CIA

Mobility for older adults with physical disabilities or cognitive impairments

LGBT older adults discriminated against in housing

Undocumented immigrants currently incarcerated and/or underpaid

Undocumented immigrants for whom prosecutorial discretion has been used to justify release or general punishment

Measurements for global web users that take into account shared devices and VPNs

Firm statistics on how often police arrest women for making false rape reports

Master database that details if/which Americans are registered to vote in multiple states

Total number of local and state police departments using stingray phone trackers (IMSI-catchers)

How much Spotify pays each of its artists per play of song

https://github.com/MimiOnuoha/missing-datasets

Race/ethnicity of actors on Broadway shows, 2014-15 season

■ White ■ Black ■ Asian ■ Hispanic ■ Other   Cast members

Show	Cast members
Constellations	2
Curious Incident ...	4
The River	5
This is Our Youth	5
Love Letters	6
Skylight	6
The Country House	6
Disgraced	9
Hand to God	9
It's Only a Play	10
Living on Love	10
A Delicate Balance	11
The Heidi Chronicles	11
The Real Thing	11
Fun Home	14
The Elephant Man	15
Airline Highway	16
It Shoulda Been You	18
The Audience	18
Gigi	22
The Visit	23
Wolf Hall	23
Honeymoon in Vegas	24
You Can't Take It With You	25
Holler If Ya Hear Me	27
Fish in the Dark	29
Something Rotten!	29
The Last Ship	29
Finding Neverland	30
Side show	30
On the Town	31
Doctor Zhivago	32
An American in Paris	33
The King and I	69

4. how **data** is visualized is **definitely** a design question

5. design for interpretation

AI System
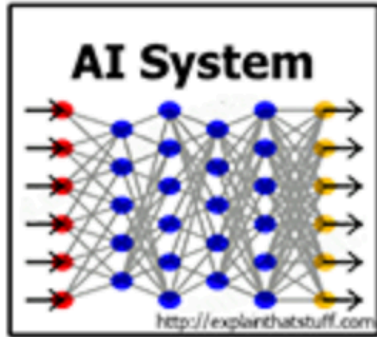
- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

DoD and non-DoD Applications

Transportation

Security

Medicine

Finance

Legal

Military

User

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

https://www.darpa.mil/program/explainable-artificial-intelligence

Design for **interpretation.**
It's not enough to be transparent.
Transparency can:

- be harmful
- intentionally confuse things
- have technical & temporal limitations

—Mike Ananny & Kate Crawford, "Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability," *new media & society* 2016.

The New York Times Magazine

Can an Algorithm Tell When Kids Are in Danger?

Child protective agencies are haunted when they fail to save kids. Pittsburgh officials believe a new data analysis program is helping them make better judgment calls.

By DAN HU...

Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women

By Reuters

Oct. 10, 2018

SAN FRANCISCO — Amazon.com Inc's machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Automation has been key to Amazon's e-commerce inside warehouses or driving pricing decisions. The experimental hiring tool used artificial intelligence candidates scores ranging from one to five stars - m rate products on Amazon, some of the people said.

"Everyone wanted this holy grail," one of the people literally wanted it to be an engine where I'm going resumes, it will spit out the top five, and we'll hire t

---

Defense Innovation Board to Explore the Ethics of AI in War

LAURENT T/SHUTTERSTOCK.COM

By Aaron Boyd,
Senior Editor

OCTOBER 11, 2018

The advisory group will provide the Pentagon with recommendations after hearing from academics, industry and the public.

ARTIFICIAL INTELLIGENCE   DEFENSE

Opinion

When Your Boss Is an Algorithm

For Uber drivers, the workplace can feel like a world of constant surveillance, automated manipulation and threats of "deactivation."

By Alex Rosenblat
Ms. Rosenblat is the author of the forthcoming book "Uberland: How Algorithms Are Rewriting the Rules of Work."

Oct. 12, 2018
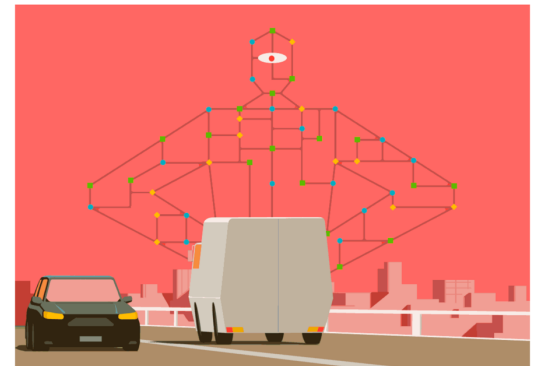
Jonathan Djob Nkondo

At least 50* different ethics...

Frameworks

Toolkits

Principles

Codes of conduct

Checklists

Oaths

Manifestos

(via Kathy Baxter, Architect of Ethical AI Practice at Salesforce, who said in a personal communication that the number may be closer to 200)

are they **proven?**

do they **work?**

UNDER SECRETARY OF DEFENSE FOR INTELLIGENCE

officium nostrum est adiuvare

ALGORITHMIC WARFARE CROSS-FUNCTIONAL TEAM

Dear Sundar,

We believe that Google should not be in the business of war. Therefore we ask that Project Maven be cancelled, and that Google draft, publicize and enforce a clear policy stating that neither Google nor its contractors will ever build warfare technology.

Google is implementing Project Maven, a customized AI surveillance engine that uses "Wide Area Motion Imagery" data captured by US Government drones to detect vehicles and other objects, track their motions, and provide results to the Department of Defense.

Recently, Googlers voiced concerns about Maven internally. Diane Greene responded, assuring them that the technology will not "operate or fly drones" and "will not be used to launch weapons." While this eliminates a narrow set of direct applications, the technology is being built for the military, and once it's delivered it could easily be used to assist in these tasks.

This plan will irreparably damage Google's brand and its ability to compete for talent. Amid growing fears of biased and weaponized AI, Google is already struggling to keep the public's trust. By entering into this contract, Google will join the ranks of companies like Palantir, Raytheon, and General Dynamics. The argument that other firms, like Microsoft and Amazon, are also participating doesn't make this any less risky for Google. Google's unique history, its motto *Don't Be Evil*, and its direct reach into the lives of billions of users set it apart.

We cannot **outsource the moral responsibility of our technologies to third parties**. Google's stated values make this clear: *Every one of our users is trusting us. Never jeopardize that. Ever.* This contract puts Google's reputation at risk and stands in direct opposition to our core values. **Building this technology to assist the US Government in military surveillance – and potentially lethal outcomes – is not acceptable.**

Recognizing Google's moral and ethical responsibility, and the threat to Google's reputation, we request that you:

1. Cancel this project immediately
2. Draft, publicize, and enforce a clear policy stating that neither Google nor its contractors will ever build warfare technology

AI

AI at Google: our principles

Sundar Pichai
CEO

Published Jun 7, 2018

At its heart, AI is computer programming that learns and adapts. It can't solve every problem, but its potential to improve our lives is profound. At Google, we use AI to make products more useful—from email that's spam-free and easier to compose, to a digital assistant you can speak to naturally, to photos that pop the fun stuff out for you to enjoy.

Beyond our products, we're using AI to help people tackle urgent problems. A pair of high school students are building AI-powered sensors to predict the risk of wildfires. Farmers are using it to monitor the health of their herds. Doctors are starting to use AI to help diagnose cancer and prevent blindness. These clear benefits are why Google invests heavily in AI research and development, and makes AI technologies widely available to others via our tools and open-source code.

We recognize that such powerful technology raises equally powerful questions about its use. How AI is developed and used will have a significant impact on society for many years to come. As a leader in AI, we feel a deep responsibility to get this right. So today, we're announcing seven principles to guide our work going forward. These are not theoretical concepts; they are concrete standards that will actively govern our research and product development and will impact our business decisions.

We acknowledge that this area is dynamic and evolving, and we will approach our work with humility, a commitment to internal and external engagement, and a willingness to adapt our approach as we learn over time.

Objectives for AI applications

We will assess AI applications in view of the following objectives. We believe that AI should:

1. Be socially beneficial.

The expanded reach of new technologies increasingly touches society as a whole. Advances in AI will have transformative impacts in a wide range of fields, including healthcare, security, energy, transportation, manufacturing, and entertainment. As we consider potential development and uses of AI technologies, we will take into account a broad range of social and economic factors, and will proceed where we believe that the overall likely benefits substantially exceed the foreseeable risks and downsides.

AI also enhances our ability to understand the meaning of content at scale. We will strive to make high-quality and accurate information readily available using AI, while continuing to respect cultural, social, and legal norms in the countries where we operate. And we will continue to thoughtfully evaluate when to make our technologies available on a non-commercial basis.

2. Avoid creating or reinforcing unfair bias.

AI algorithms and datasets can reflect, reinforce, or reduce unfair biases. We recognize that distinguishing fair from unfair biases is not always simple, and differs across cultures and societies. We will seek to avoid unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief.

3. Be built and tested for safety.

We will continue to develop and apply strong safety and security practices to avoid unintended results that create risks of harm. We will design our AI systems to be appropriately cautious, and seek to develop them in accordance with best practices in AI safety research. In appropriate cases, we will test AI technologies in constrained environments and monitor their operation after deployment.

4. Be accountable to people.

We will design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal. Our AI technologies will be subject to appropriate human direction and control.

5. Incorporate privacy design principles.

We will incorporate our privacy principles in the development and use of our AI technologies. We will give opportunity for notice and consent, encourage architectures with privacy safeguards, and provide appropriate transparency and control over the use of data.

6. Uphold high standards of scientific excellence.

Technological innovation is rooted in the scientific method and a commitment to open inquiry, intellectual rigor, integrity, and collaboration. AI tools have the potential to unlock new realms of scientific research and knowledge in critical domains like biology, chemistry, medicine, and environmental sciences. We aspire to high standards of scientific excellence as we work to progress AI development.

We will work with a range of stakeholders to promote thoughtful leadership in this area, drawing on scientifically rigorous and multidisciplinary approaches. And we will responsibly share AI knowledge by publishing educational materials, best practices, and research that enable more people to develop useful AI applications.

7. Be made available for uses that accord with these principles.

Many technologies have multiple uses. We will work to limit potentially harmful or abusive applications. As we develop and deploy AI technologies, we will evaluate likely uses in light of the following factors:

- *Primary purpose and use*: the primary purpose and likely use of a technology and application, including how closely the solution is related to or adaptable to a harmful use

- *Nature and uniqueness*: whether we are making available technology that is unique or more generally available

- *Scale*: whether the use of this technology will have significant impact

- *Nature of Google's involvement*: whether we are providing general-purpose tools, integrating tools for customers, or developing custom solutions

AI applications we will not pursue

In addition to the above objectives, we will not design or deploy AI in the following application areas:

1. Technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints.

2. Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.

3. Technologies that gather or use information for surveillance violating internationally accepted norms.

4. Technologies whose purpose contravenes widely accepted principles of international law and human rights.

We want to be clear that while we are not developing AI for use in weapons, we will continue our work with governments and the military in many other areas. These include cybersecurity, training, military recruitment, veterans' healthcare, and search and rescue. These collaborations are important and we'll actively look for more ways to augment the critical work of these organizations and keep service members and civilians safe.

AI for the long term

While this is how we're choosing to approach AI, we understand there is room for many voices in this conversation. As AI technologies progress, we'll work with a range of stakeholders to promote thoughtful leadership in this area, drawing on scientifically rigorous and multidisciplinary approaches. And we will continue to share what we've learned to improve AI technologies and practices.

We believe these principles are the right foundation for our company and the future development of AI. This approach is consistent with the values laid out in our original Founders' Letter back in 2004. There we made clear our intention to take a long-term perspective, even if it means making short-term tradeoffs. We said it then, and we believe it now.

ethics can act as "**window-dressing**" or a "**mask**" for the misdeeds of companies.

Chief Ethics Officers: Who Needs Them?

By Hannah Clark

Having a chief ethics officer didn't help Hewlett-Packard . You've heard the story: Chairwoman **Patricia Dunn** lost her job after hiring private investigators to find leakers on HP's board. The spying scandal that ensued led to Dunn's indictment and an investigation by the House Energy and Commerce Committee. Even CEO Mark Hurd Mark Hurd , who replaced Dunn as chair, has been implicated in the scandal. And it all happened under the watch of **Kevin Hunsaker**, HP's senior counsel and chief ethics officer. He resigned in September.

Chief ethics and compliance officers have become trendy in recent years, but some experts fear they act mainly as window dressing. If one person is in charge of ethics, they argue, everyone else might think they're off the hook. "In a way, it's a job creation program," says Mary Ann Jorgenson, a partner in the law firm Squire, Sanders & Dempsey. "It's not great for every company. It's not necessary for every company."
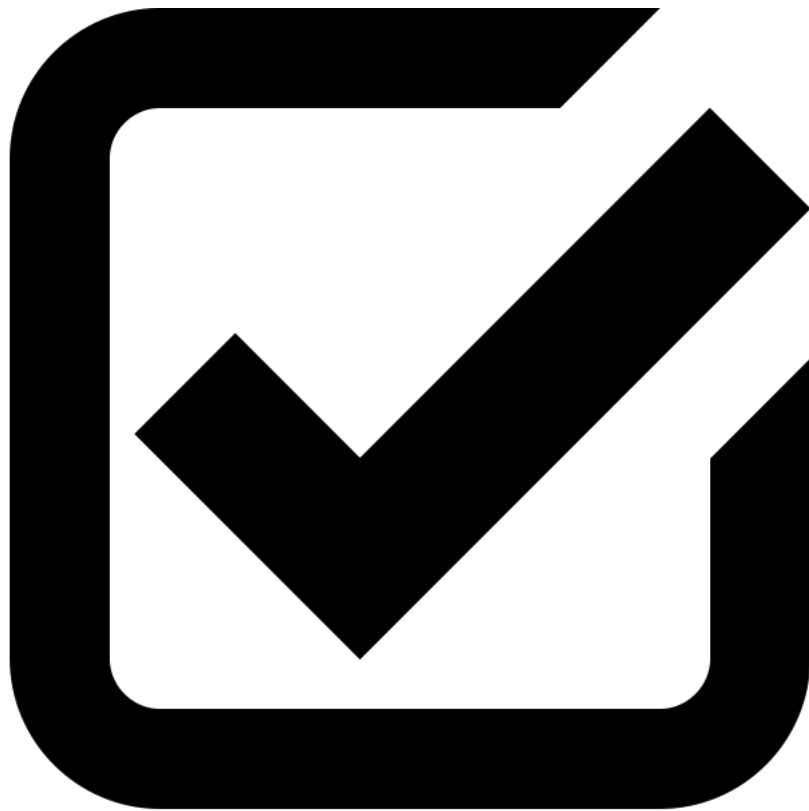
"It's an 'A for effort' philosophy, in which companies that prioritize ethics can sometimes escape punishment when their ethics programs fail."

—Hannah Clark, *Fortune,* 2006

ethics as a hedge

ethics as a hedge against **litigation**

ethics as a hedge against regulation

ethics-washing | ethics-shopping

what happens in 2021 when ethics is no longer a **hot button** topic?

A two-year-old's solution to the trolley problem (almost 14 million views!)
E.J. Masicampo, https://www.youtube.com/watch?v=-N_RZJUAQY4

what do we *really* mean when we say ethics?

thank you!

girlwonder.com | @maximolly