

Artificial Intelligence Ethics

Sven Koenig, USC

Russell and Norvig, 3rd Edition, Section 26.3

These slides are new and can contain mistakes and typos.
Please report them to Sven (skoenig@usc.edu).

1

Consumer Products



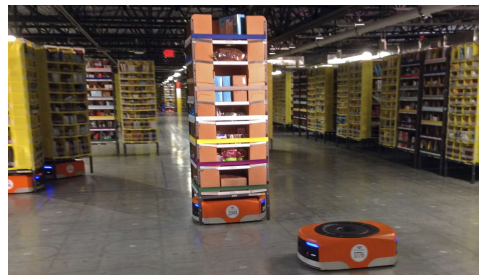
2

Amazon Fulfillment Centers

- 2003 Kiva Systems founded
- 2012 Amazon acquires Kiva for \$775 million
- 2015 Kiva Systems becomes Amazon Robotics



[www.npr.org – Getty Images]



[www.theguardian.com - AP]



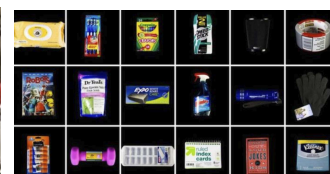
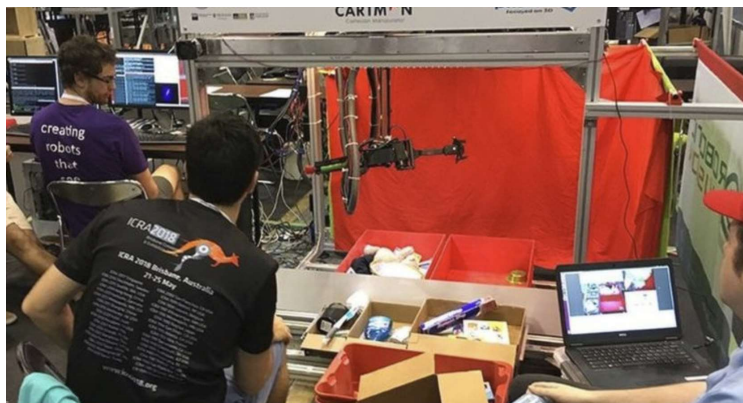
ICAPS 2014
Invited talk by
Peter Wurman
How to Coordinate
a Thousand Robots



- > 3,000 robots on > 110,000 square meters in Tracy, California

3

Amazon Picking/Robotics Challenge



4

DARPA Robotics Challenge

- 2015



[youtube.com]

- “If you are worried about the TERMINATOR, just keep your door closed.”

5

Game Playing: Go (Google Deepmind)

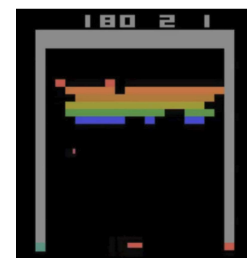
- 2016



[Go Game Guru]

AlphaGo vs. Lee Sedol

4-1



[PC World]

6

Science Tests

- 2016

Intelligent Machines

The Best AI Program Still Flunks an Eighth-Grade Science Test

**MIT
Technology
Review**

A contest designed to push the limits of artificial intelligence suggests that truly intelligent machines are a long way off.

by Will Knight February 17, 2016

A score of almost 60 percent might disappoint most parents, but it is remarkable for a computer. The test used for the contest was, however, simplified slightly to make it practical for computers to attempt. Diagrams were removed, for example, and only questions with multiple-choice answers were used.

7

Loebner Competition (Turing Test)

1950

"Computing Machinery and Intelligence"
- a paper by Alan Turing



- 2017

Hello, my name is Andrew. What's your name?
Hello there. Your name is Andrew. My name is Millie.

Why don't you tell me a little more about yourself?

My name is Millie and I am a human volunteer for the Loebner Prize. I live in Leeds and am single. I am interested in AI but don't have a chatbot myself.

What about you?

I am a researcher in Artificial Intelligence at Goldsmiths University, do you know what that is?
Do you enjoy your research?

I am a volunteer who has agreed to help with a Turing test.

Will you tell me about your dreams?
I dream that I will pass my university exams.

Can you rephrase that?
No, not really. What don't you understand?

What is my occupation?
Well you said your job was RESEARCHER.

Can you tell me about a film you haven't seen?
you have not seen = Are you asking about my have not seen?

What do you think of Trump?
I read that over 50% were either concerned or scared about Donald Trump being President. if that's the case, why did they vote for him?

Anything else?
That is all I have for right now.

What is the answer to "Add 34957 to 70764"?
34957 to 70764 = Ah yes, 34957 is a type of cheese that is made from sheep's milk. It's not very popular though.

How do you recommend I make tea?
I suppose I would recommend something if I'd had a pleasant experience with it in the past.

What do you get if you bake dough?
Is this a joke? I don't know Andrew. What do you get?

8

Game Playing: Soccer

- 2018



9

Some are concerned...

the guardian

US politics world opinion sports soccer tech arts lifestyle fashion business travel environment

home > tech

Artificial intelligence (AI) **Elon Musk: artificial intelligence is our biggest existential threat**

The AI investor says that humanity risks 'summoning a demon' and calls for more regulatory oversight

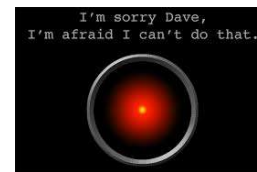
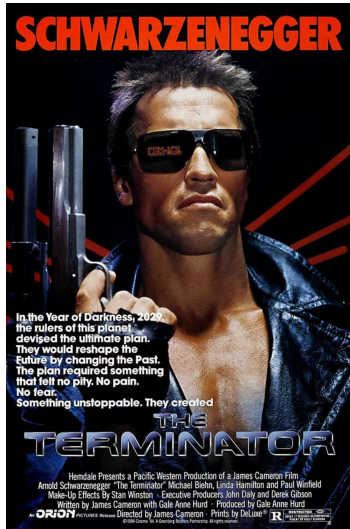
TECH TRANSFORMERS | A CNBC SPECIAL REPORT

TECH | MOBILE | SOCIAL MEDIA | ENTERPRISE | CYBERSECURITY | TECH GUIDE

Stephen Hawking says A.I. could be 'worst event in the history of our civilization'

10

Movies paint a dark picture...



11

State of the Art in Intelligent Systems (= Agents)

- Areas of artificial intelligence
 - Knowledge Representation and Reasoning
 - Planning
 - Machine Learning
 - Multi-agent coordination
 - ...
 - Robotics
 - Vision
 - Natural language processing
 - ...



12

State of the Art in Intelligent Systems (= Agents)

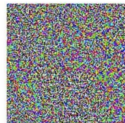
- Headlines in the news

- 2012: A Massive Google Network Learns to Identify Cats [npr.org]

- 2015:



+ .007 ×



=



"panda"
57.7% confidence

"nematode"
8.2% confidence

"gibbon"
99.3 % confidence

[popsci.com]



[hexus.com]

- 2017: This Google AI Built to Identify Cat Pics Can Recognize Gene Mutations [popularmechanics.com]
- 2018: Google Lens Can Now Identify Dog and Cat Breeds [fortune.com] "The new breed identification skill seems to work well for purebred dogs, but is more hit or miss for mixed breed dogs..."

13

State of the Art in Intelligent Systems (= Agents)

- Limitations [Marcus 2017]

- Needs lots of data
- Limited capacity for transfer
- Struggles with open-ended inference
- Not sufficiently transparent
- Not sufficiently integrated with prior knowledge
- Does not sufficiently distinguish causation from correlation
- Presumes largely a stable world
- Answers cannot be sufficiently trusted
- Difficult to engineer with

14

State of the Art in Intelligent Systems (= Agents)

- Autonomous agents
 - Rational agents (= agents that make good decisions)
 - Narrowly intelligent systems (task level)
 - Single AI technique

15

State of the Art in Intelligent Systems (= Agents)

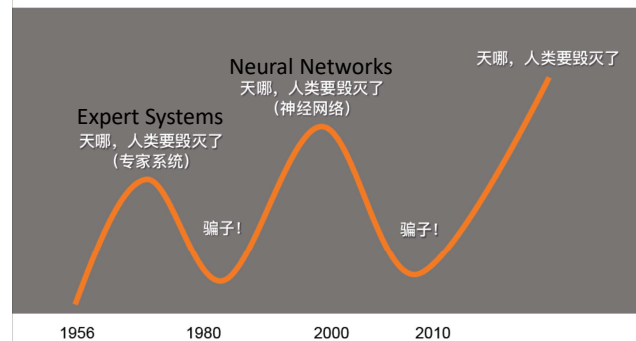
- Autonomous agents
 - Rational agents (= agents that make good decisions)
 - Narrowly intelligent systems (task level)
 - Single AI technique
 - Broadly intelligent systems (job level)
 - Integration of AI techniques
 - Believable agents (= agents that behave like humans)
 - Human-aware agents with human-like interactions via gestures, speech, ...
 - Agents that can understand and imitate emotions
 - Cognitive agents (= agents that think like humans)

16

Artificial Intelligence in 2028

- Kai-Fu Lee (Sinovation Ventures; Founder and Managing Director of Microsoft Research Asia, China 1998-2000)

Waves of Artificial Intelligence since 1956
人工智能自1956年来的几波浪潮

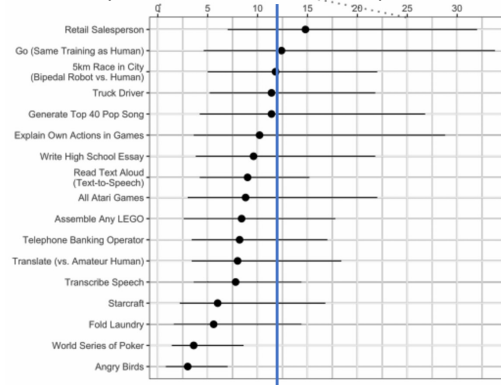


17

Artificial Intelligence in 2028

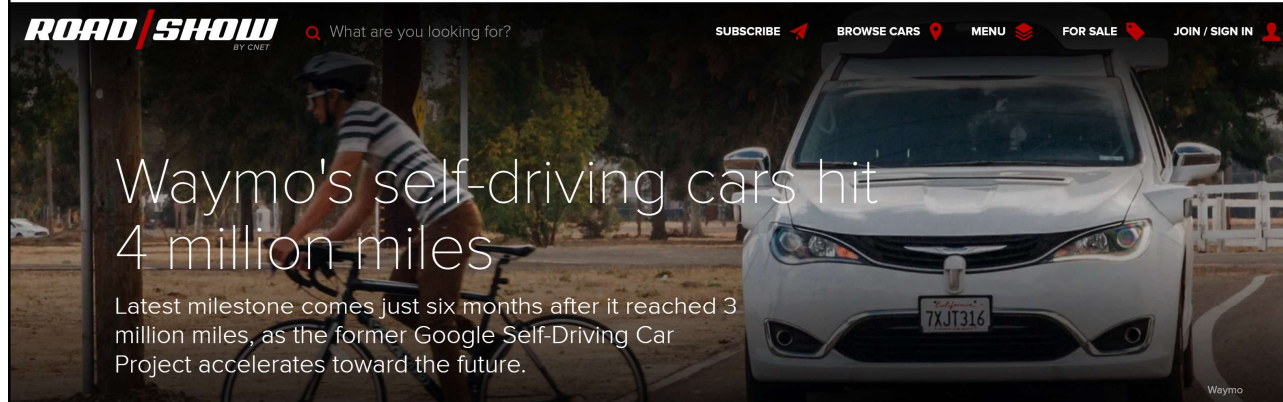
- AI and Life in 2030 – One hundred year study on AI
- Survey of the Future of Humanity Institute of the University of Oxford

years from 2016 until AI outperforms humans



18

Self-Driving Cars



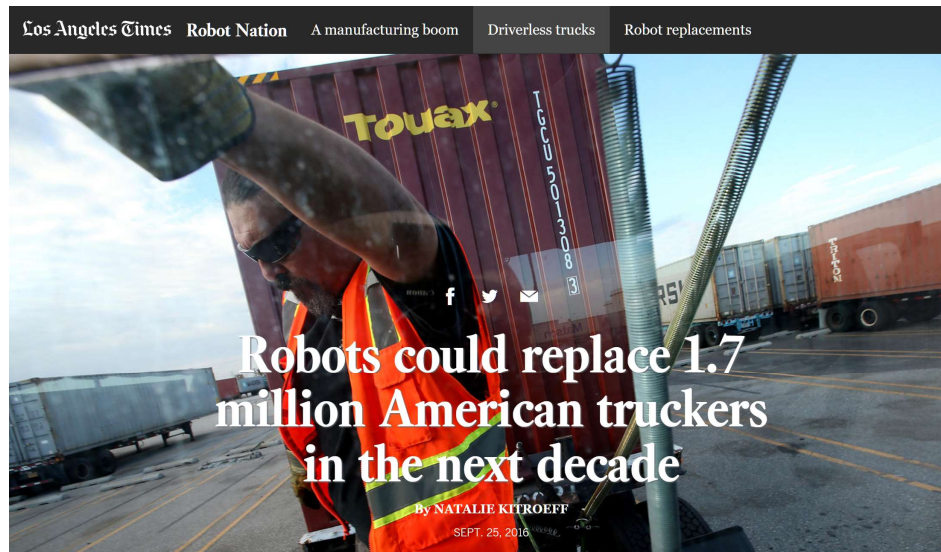
19

Self-Driving Cars

- Imagine that you are on the design team of a self-driving car. Should you worry about the following issue facing the planning system:
- The car notices that it made a mistake and is driving at full speed toward a kid on the street. It has only two options:
 - Keep going straight (and break), which kills the kid.
 - Turn away from the kid (and break), which crashes the car into a wall and kills the driver.
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

20

Self-Driving Cars



21

Self-Driving Cars



[JPL/NASA]

22

Self-Driving Cars

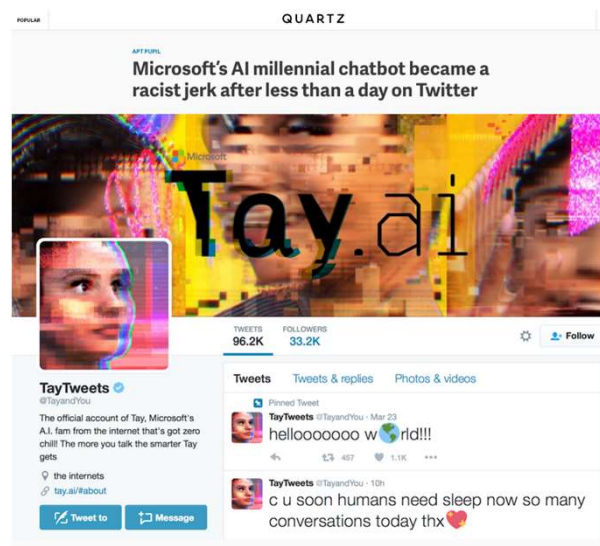


[JPL/NASA]

The women of the Jet Propulsion Laboratory helped launch the first American satellites, lunar missions and planetary explorations. Those "human computers," as they were called, are seen here in 1953.

23

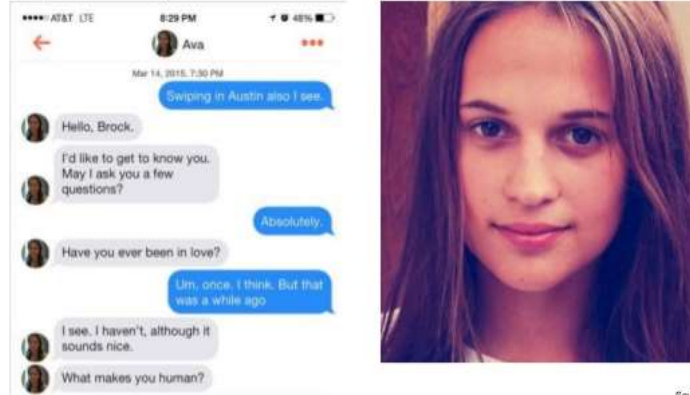
Chatbots



24

Chatbots

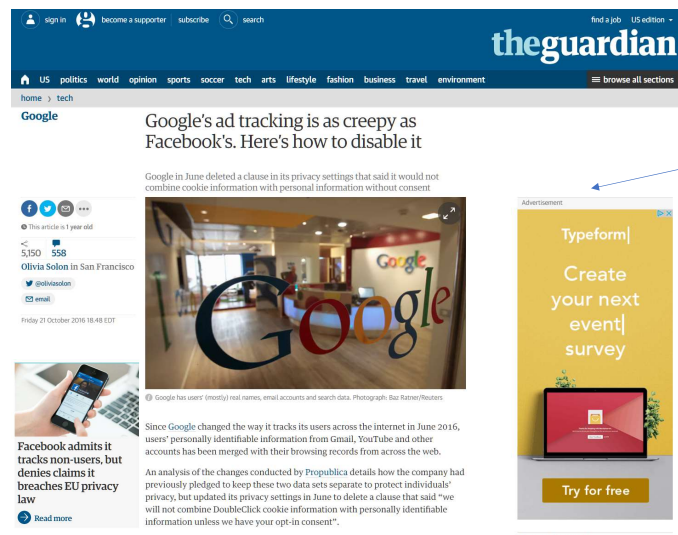
Tinder bot preying on Austin dating scene actually a bot promoting movie "Ex Machina"



Source

25

Targeted Advertising



targeted advertising

26

Targeted Advertising

- Imagine that you are on the design team of a system that selects targeted advertising for webpages. Should you worry about the following issue facing the machine learning system:

July 07, 2015

Questioning the Fairness of Targeting Ads Online

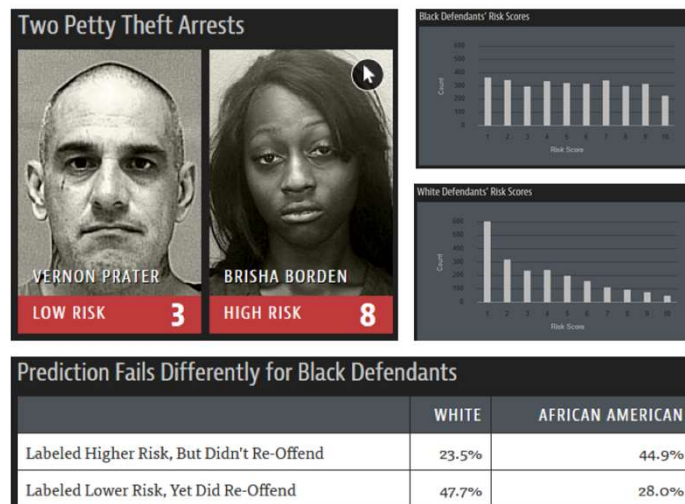
CMU Probes Online Ad Ecosystem

By [Byron Spice](#) / 412-268-9068

Experiments by Carnegie Mellon University showed that significantly fewer women than men were shown online ads promising them help getting jobs paying more than \$200,000, raising questions about the fairness of targeting ads online.

27

Decision-Support Systems



28

Issues

- AI systems can process large quantities of data, detect regularities in them, draw inferences from them and determine effective courses of action - sometimes faster and better than humans and sometimes as part of hardware that is able to perform many different, versatile and potentially dangerous actions.
- The behavior of AI systems can be difficult to validate, predict or explain since they are complex, reason in ways different from humans and can change their behavior via learning.
- Their behavior can also be difficult to monitor by humans in case of fast decisions, such as buy and sell decisions on stock markets.

29

Issues

- Do we need to worry about the reliability, robustness, and safety of AI systems?
- Do we need to provide oversight of their operation?
- How do we guarantee that their behavior is consistent with social norms and human values?
- Who is liable for incorrect AI decisions?
- How will AI technology impact standard of living, distribution and quality of work, and other social and economic aspects?

30

Issues

- Top 10 ethical issues in AI according to the World Economic Forum
 1. **Unemployment.** What happens after the end of jobs?
 2. **Inequality.** How do we distribute the wealth created by machines?
 3. **Humanity.** How do machines affect our behavior and interaction?
 4. **Artificial stupidity.** How can we guard against mistakes?
 5. **Racist robots.** How do we eliminate AI bias?
 6. **Security.** How do we keep AI safe from adversaries?
 7. **Evil genies.** How do we protect against unintended consequences?
 8. **Singularity.** How do we stay in control of a complex intelligent system?
 9. **Robot rights.** How do we define the humane treatment of robots?

31

Issues

- Should AI systems be allowed to pretend to be human?
- More generally, should AI systems be allowed to lie?
- Should autonomous weapons be banned, just like the UN banned blinding laser weapons?
- ...

32

Ethics

- A branch of philosophy that involves systematizing, defending, and recommending concepts of right and wrong conduct
- Seeks to resolve questions of human morality by defining concepts such as good and evil, right and wrong, virtue and vice, justice and crime
- Normative ethics studies how to determine a moral course of action

33

Law-Based Ethics (Deontology)

- Example: Immanuel Kant
- Questions: What is my duty? What are the right rules (= universal moral law) to follow?
- Issue: How do we apply these rules to decision situations?
- How would we implement this with tools learned in CS360?

34

Law-Based Ethics (Deontology)



The Ten Commandments [\[change | change source \]](#)

The Ten Commandments, also known as the Decalogue, found in the Ark of the Covenant are:

1. I am the Lord your God, You shall have no other gods before me.
2. You shall not make for yourself an image in the form of anything in heaven above or on the earth beneath or in the waters below. You shall not bow down to them or worship them.
3. You shall not misuse the name of the Lord your God.
4. Remember the Sabbath day by keeping it holy.
5. Honour your father and your mother.
6. You shall not murder.
7. You shall not commit adultery.
8. You shall not steal.
9. You shall not give false testimony against your neighbour,
10. You shall not covet anything that belongs to your neighbour.

35

Law-Based Ethics (Deontology)

- Isaac Asimov's three laws of robotics
 - A robot may not injure a human being or, through inaction, allow a human being to come to harm.
 - A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
 - A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

36

Utilitarian Ethics (Consequentialism)

- Example: Jeremy Bentham and John Stuart Mill
- Questions: What is the greatest possible good for the greatest number? What does a cost-benefit analysis recommend?
- Issues: How to define and measure goodness? How to weight goodness for different individuals?
- How would we implement this with tools learned in CS360?

37

Virtue Ethics (Teleological Ethics)

- Example: Aristotle
- Questions: Who should I be? What's the best behavior in this particular situation? How to develop habits and dispositions that help people to achieve their goals and help them flourish as an individual?
- Issue: How do we make virtue ethics operational?

38

Common Sense Morality

- Resnik's eight principles (norms, not laws)
 - **Non-malificence:** Do not harm yourself or other people.
 - **Beneficence:** Help yourself and other people.
 - **Autonomy:** Allow rational individuals to make free and informed choices.
 - **Justice:** Treat people fairly: treat equals equally, unequals unequally.
 - **Utility:** Maximize the ratio of benefits to harms for all people.
 - **Fidelity:** Keep your promises and agreements.
 - **Honesty:** Do not lie, defraud, deceive or mislead.
 - **Privacy:** Respect personal privacy and confidentiality.

39

Targeted Advertising

The screenshot shows the Guardian website interface. At the top, there's a navigation bar with links like 'sign in', 'become a supporter', 'subscribe', and 'search'. Below this is a secondary navigation bar with categories like 'US', 'politics', 'world', 'opinion', 'sports', 'tech', 'arts', 'lifestyle', 'fashion', 'business', 'travel', and 'environment'. The main content area features an article titled 'Google's ad tracking is as creepy as Facebook's. Here's how to disable it' by Olivia Solon. The article text discusses Google's privacy changes in June 2016, where it deleted a clause that would not combine DoubleClick cookie information with personal information without consent. A large image of the Google logo is part of the article. To the right of the article is a yellow advertisement for Typeform, which says 'Create your next event survey' and 'Try for free'.

40

Targeted Advertising

- **Law-Based Ethics**
 - A consideration might be that the collection of user data is only permissible with the explicit consent of the user.
- **Utilitarian Ethics**
 - Considerations might be the need for revenue for the provider of free web services, the utility the user might derive from discovering new opportunities, and the user's discomfort of having their data shared.
- **Virtue Ethics**
 - A consideration might be that the user should concentrate on their work, not ads.

41

Ethical agents

- James H. Moor defines four types of ethical agents
 1. **Ethical impact agents** are agents whose actions have ethical consequences whether intended or not. (Example: Knife.)
 2. **Implicit ethical agents** have ethical considerations hardcoded into their design. (Example: Seat belt.)
 3. **Explicit ethical agents** can reason about ethics, that is, identify and process ethical information about a variety of situations and make sensitive determinations about what should be done.
 4. **Full ethical agents** make explicit moral judgments about a wide variety of situations and justify them.

42

Initiatives

- Partnership on AI
 - <https://www.partnershiponai.org>
- IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems
 - https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
- AI Now Institute
 - <https://ainowinstitute.org>
- DARPA Explainable AI Program
 - <https://www.darpa.mil/program/explainable-artificial-intelligence>
- Algorithm Watch
 - <https://algorithmwatch.org>
- Pervasive Data Ethics
 - <https://pervade.umd.edu>
- Future of Life Institute
 - <https://futureoflife.org/>