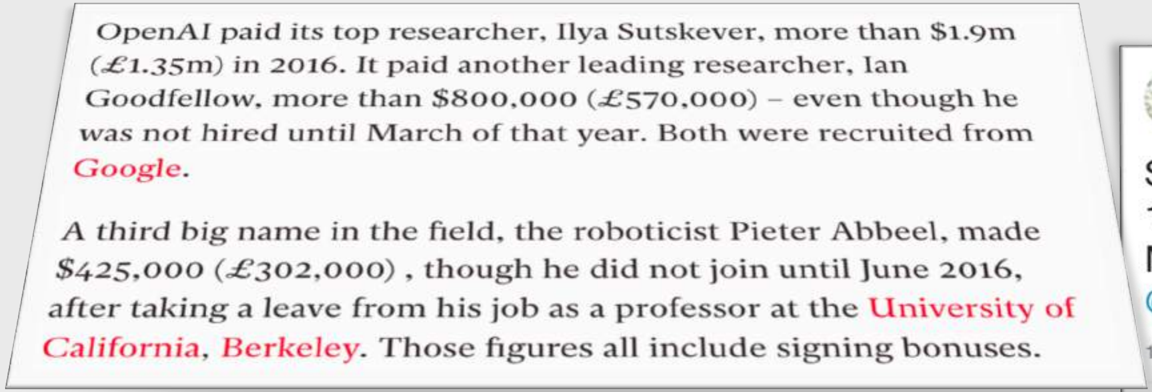
AI Ethics for Enterprise AI

Francesca Rossi

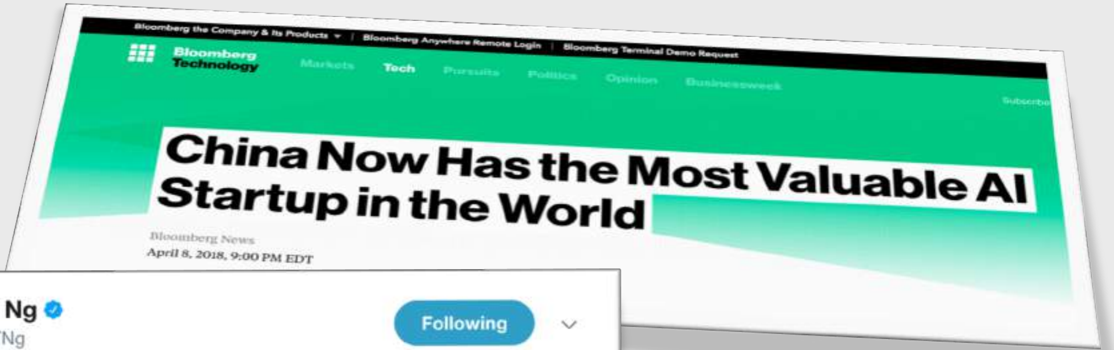
IBM AI Ethics Global Leader

AI is all the rave



OpenAI paid its top researcher, Ilya Sutskever, more than \$1.9m (£1.35m) in 2016. It paid another leading researcher, Ian Goodfellow, more than \$800,000 (£570,000) – even though he was not hired until March of that year. Both were recruited from Google.

A third big name in the field, the roboticist Pieter Abbeel, made \$425,000 (£302,000) , though he did not join until June 2016, after taking a leave from his job as a professor at the University of California, Berkeley. Those figures all include signing bonuses.

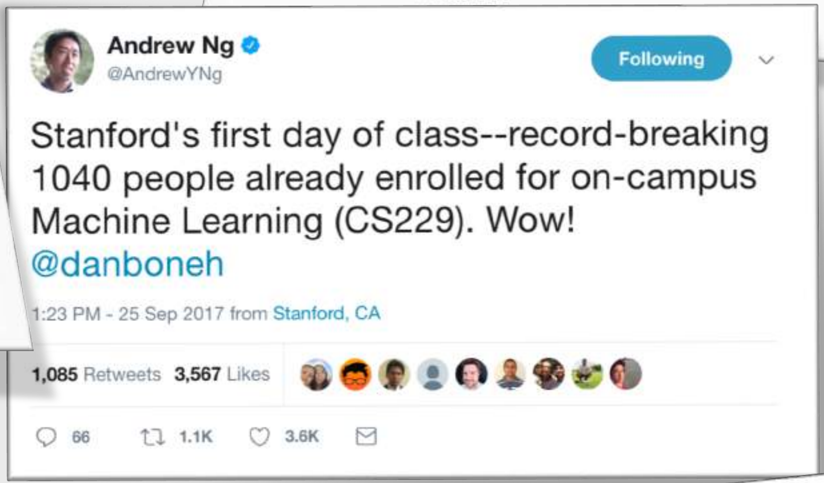


Bloomberg Technology

China Now Has the Most Valuable AI Startup in the World

Bloomberg News

April 8, 2018, 9:00 PM EDT



Andrew Ng
@AndrewYNg

Following

Stanford's first day of class--record-breaking 1040 people already enrolled for on-campus Machine Learning (CS229). Wow! @danboneh

1:23 PM - 25 Sep 2017 from Stanford, CA

1,085 Retweets 3,567 Likes

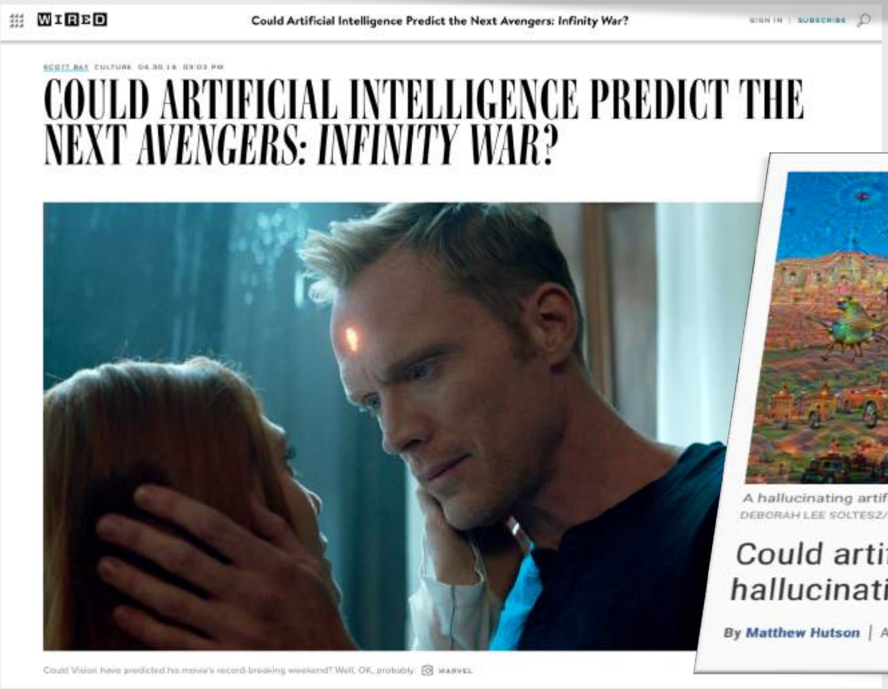
66 1.1K 3.6K

MIT Intro to Machine Learning course:

2013 – 138 students,

2016 – 302 students

2017 – 700 students



WIRED

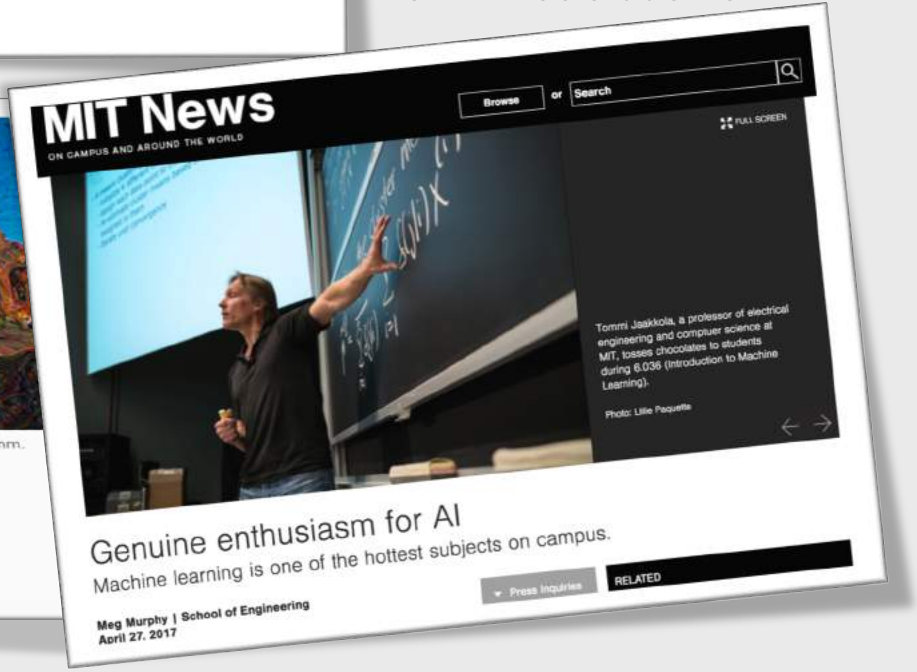
Could Artificial Intelligence Predict the Next Avengers: Infinity War?

COULD ARTIFICIAL INTELLIGENCE PREDICT THE NEXT AVENGERS: INFINITY WAR?

A hallucinating artificial intelligence might see something like this product of Google's Deep Dream algorithm.
DEBORAH LEE SOLTESZ/FLICKR

Could artificial intelligence get depressed and have hallucinations?

By Matthew Hutson | Apr. 9, 2018 , 12:10 PM



MIT News
ON CAMPUS AND AROUND THE WORLD

Tommi Jaakkola, a professor of electrical engineering and computer science at MIT, tosses chocolates to students during 6.036 (Introduction to Machine Learning).

Photo: Lillie Paquette

Genuine enthusiasm for AI

Machine learning is one of the hottest subjects on campus.

Meg Murphy | School of Engineering
April 27, 2017

In 2018, blended AI will disrupt your customer service and sales strategy

FORRESTER

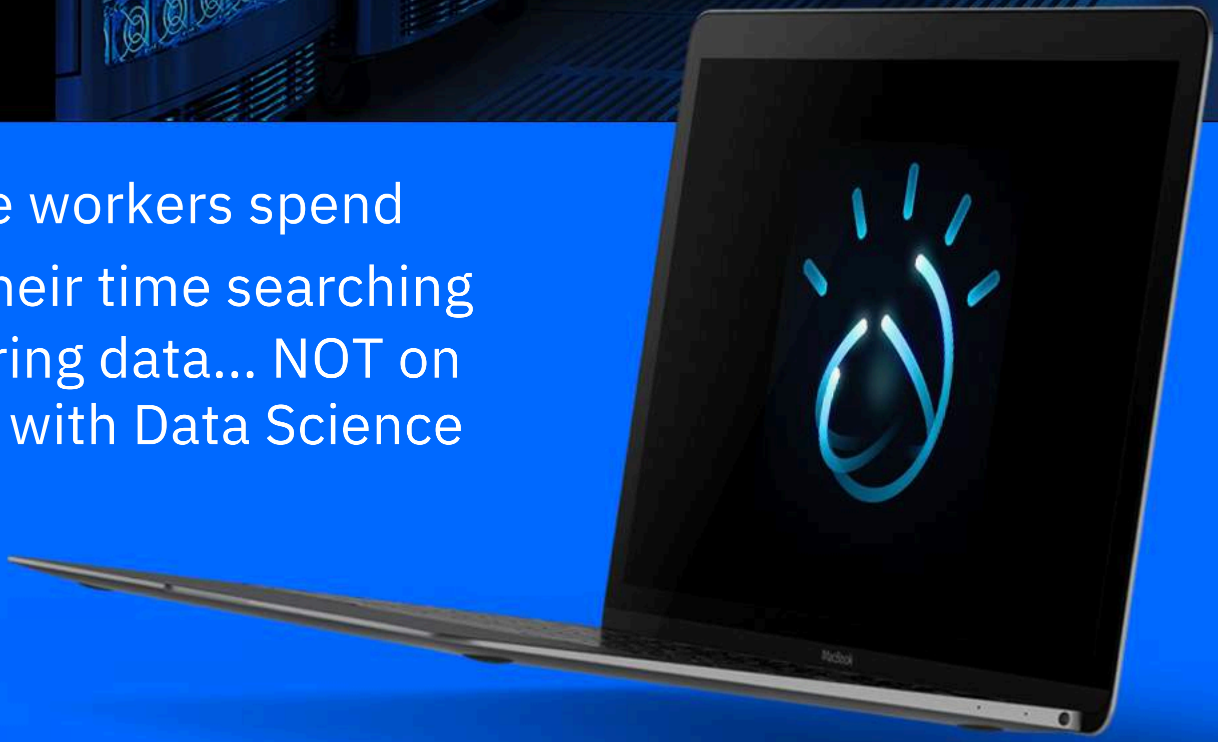
85% of CIOs will be piloting AI programs by 2020

Gartner

75% of commercial enterprise apps will use AI by 2020

IDC
Analyze the Future

Knowledge workers spend **80%** of their time searching and preparing data... NOT on innovating with Data Science and AI



Applications of AI/ML today

Home assistants (Alexa)
Travel assistants (Waze)
Ride-sharing apps (Uber, Lyft)
Auto-pilot

Client service chatbots

Friend recommendations (Facebook)
Purchase recommendations (Amazon)
Movie recommendations (Netflix)
Add placement (Google)
News curation

Medical image analysis
Treatment plan recommendation

Credit risk scoring
Loan approval
Fraud detection
Resume prioritization

Recidivism prediction (Compas)

Example of AI challenges we are tackling

Compliance

Is my organization compliant with latest regulatory documents?

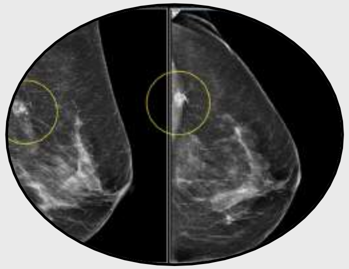
Marketing / Business

Summarize the strategic intent of a company based on recent news articles

Customer Care

Bot that can guide a user through buying the right insurance policy

Healthcare



Improve the accuracy of breast cancer screening

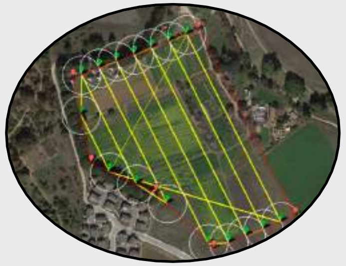
Media

Create highlights of sports events

Visual Inspection

Find rust on electric towers, using drones

IoT



Predict yield of field based on images and sensor data

Industrial

Guide me through fixing malfunctioning components

Intelligence, AI, AGI

- **Intelligence**: ability to achieve goals in a wide range of environments
- **AI -- Artificial Intelligence**: intelligence in an artificial agent
- **Current AI**: super-human capabilities in narrow domains and use cases
  - **Narrow AI**
- **AGI– Artificial General Intelligence**: An intellect that is smarter than the best human brains in practically every field, including scientific creativity, general wisdom, and social skills
  - Breath, generality, well-roundedness, versatility
  - Deep understanding, not just capability, wisdom

AI, ML, DL

ARTIFICIAL INTELLIGENCE

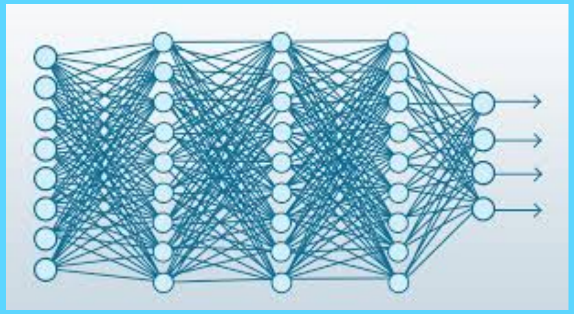
Making of intelligent
machines and programs

MACHINE LEARNING

Ability to learn without being
explicitly programmed

DEEP LEARNING

Learning based on Deep Neural
Networks

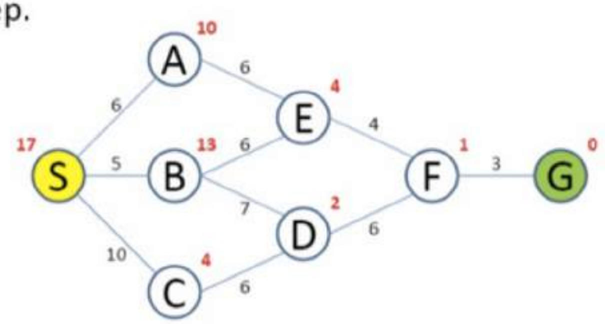


1950's | 1960's | 1970's | 1980's | 1990's | 2000's | 2006's | 2010's | 2012's | 2017's

AI: (Symbolic) Reasoning

- Exploit the knowledge to estimate the best action to take
- Not always probabilistic
- Pros:
  - Causality
  - Optimality
  - Explainability
  - Algorithm verification
- Cons:
  - Needs precise specification of the problem and solution method
  - Not suitable for ill-defined tasks

step.

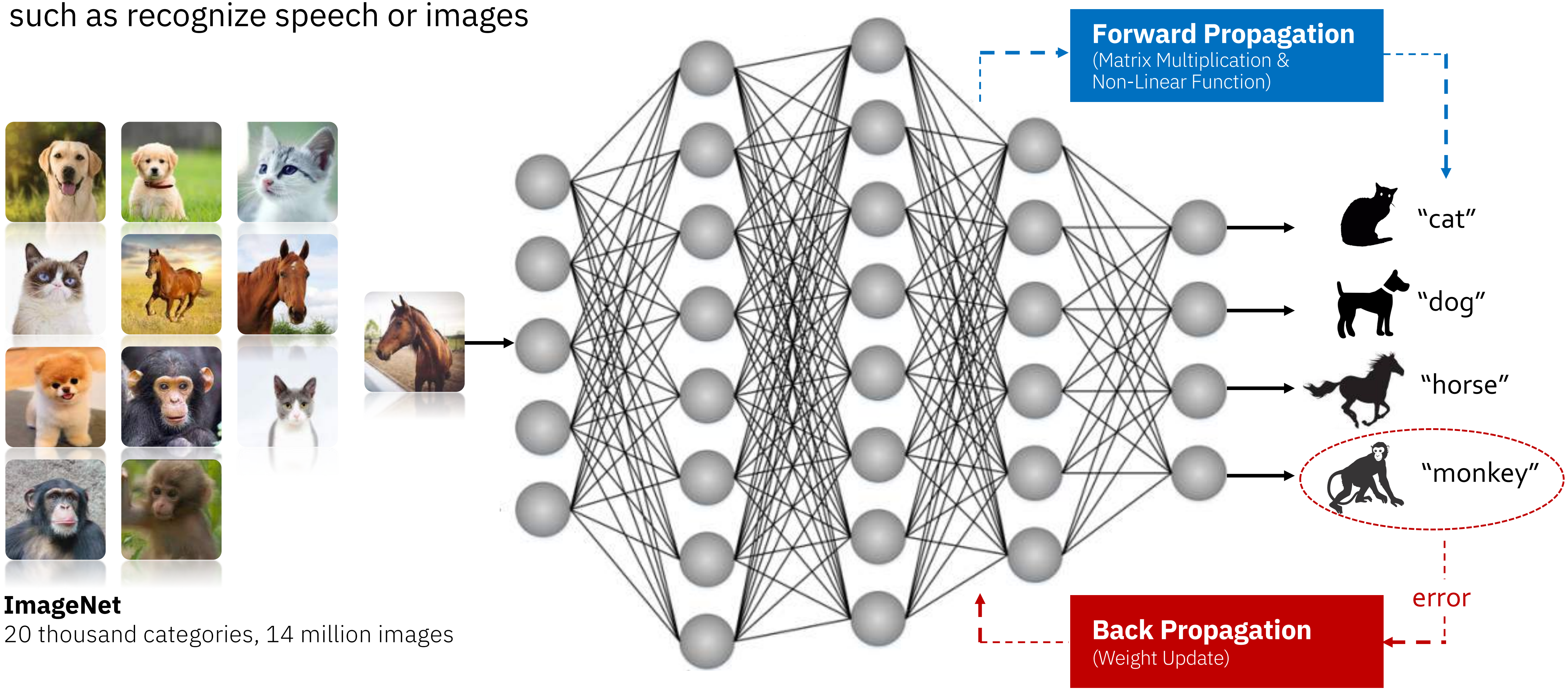


A* Algorithm

1. **Initialize**: set OPEN=[s], CLOSED=[], g(s)=0, f(s)=h(s)
2. **Fail:** If OPEN=[], then terminate and fail
3. **Select:** Select a state with minimum cost ,n, from OPEN and save in CLOSED
4. **Terminate**: If n∈G then terminate with success and return f(s)
5. **Expand:** For each successors , m of n
   For each successor, m, insert m in OPEN only if
   if m∉ [OPEN∪CLOSED]
   set g(m)= g[n]+C[n,m]
   Set f(m)=g(m)+h(n)
   if m∈[OPEN∪CLOSED]
   set g(m)= min{ g[m], g(n)+C[n,m]}
   Set f(m)=g(m)+h(m)
   If f[m] has decreased and m ∈ CLOSED move m to OPEN
6. **Loop**: Goto step 2

AI: Machine Learning

When presented with sample data, an artificial neural network can be trained to perform a specific task, such as recognize speech or images



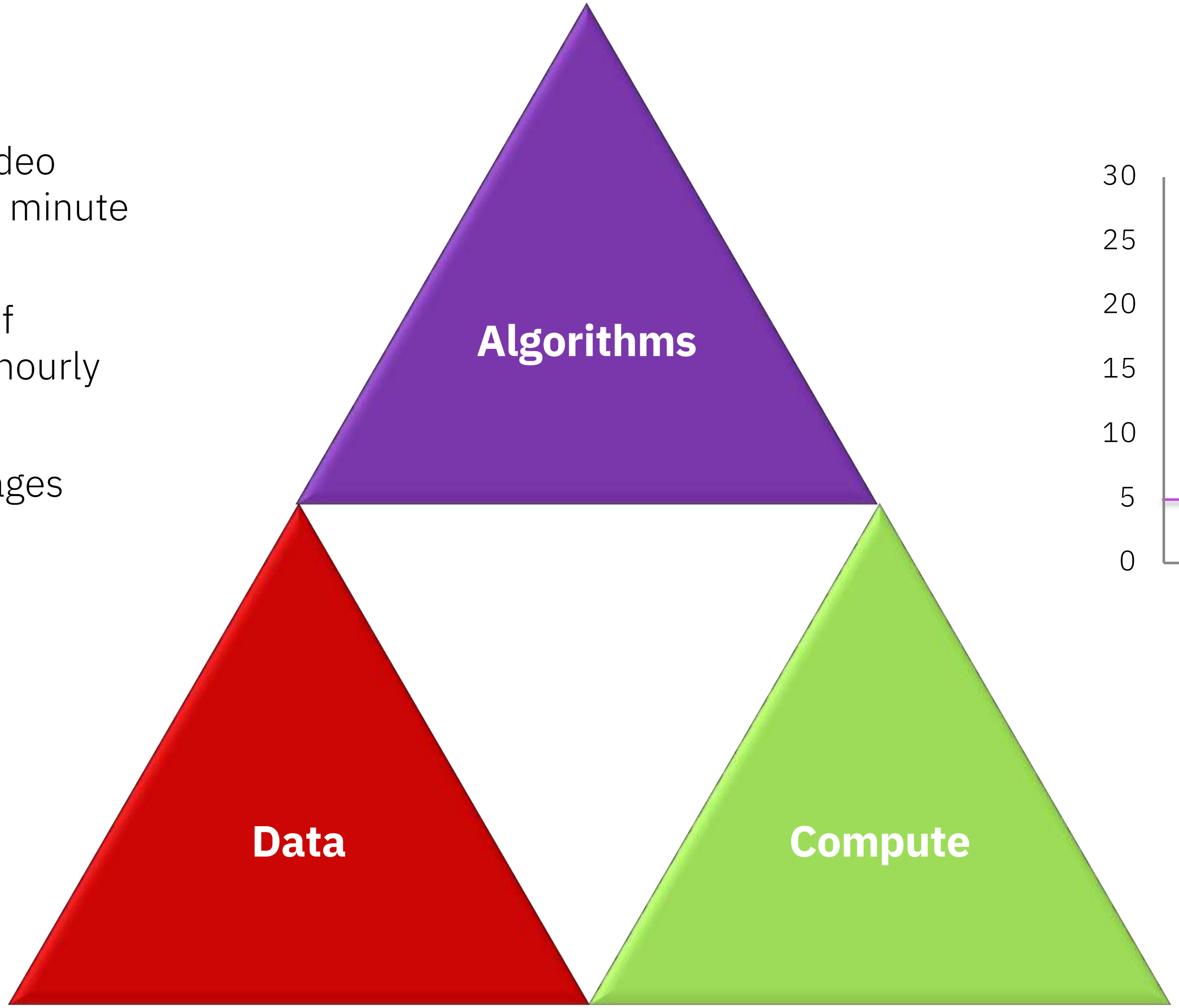
ImageNet
20 thousand categories, 14 million images

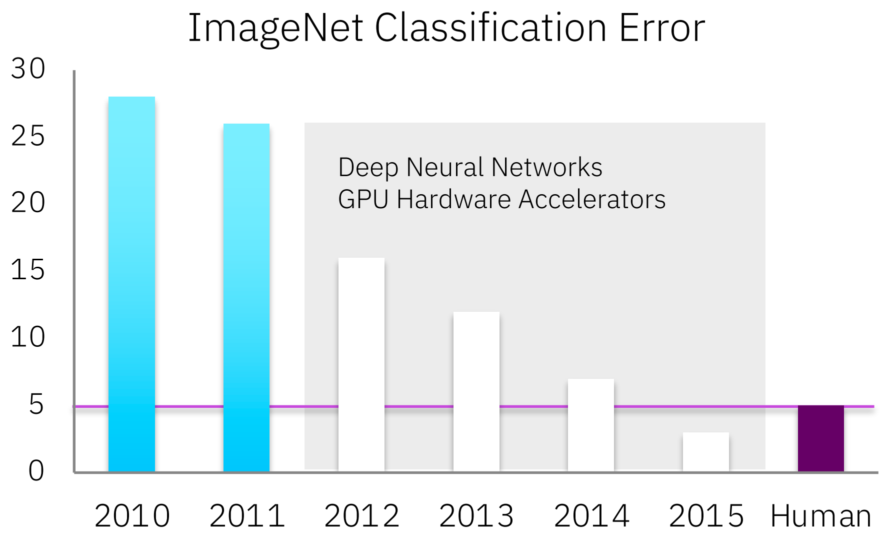
Deep Learning explosion

YouTube
400 hours of video uploaded every minute

Walmart
2.5 petabytes of customer data hourly

Facebook
350 million images uploaded daily



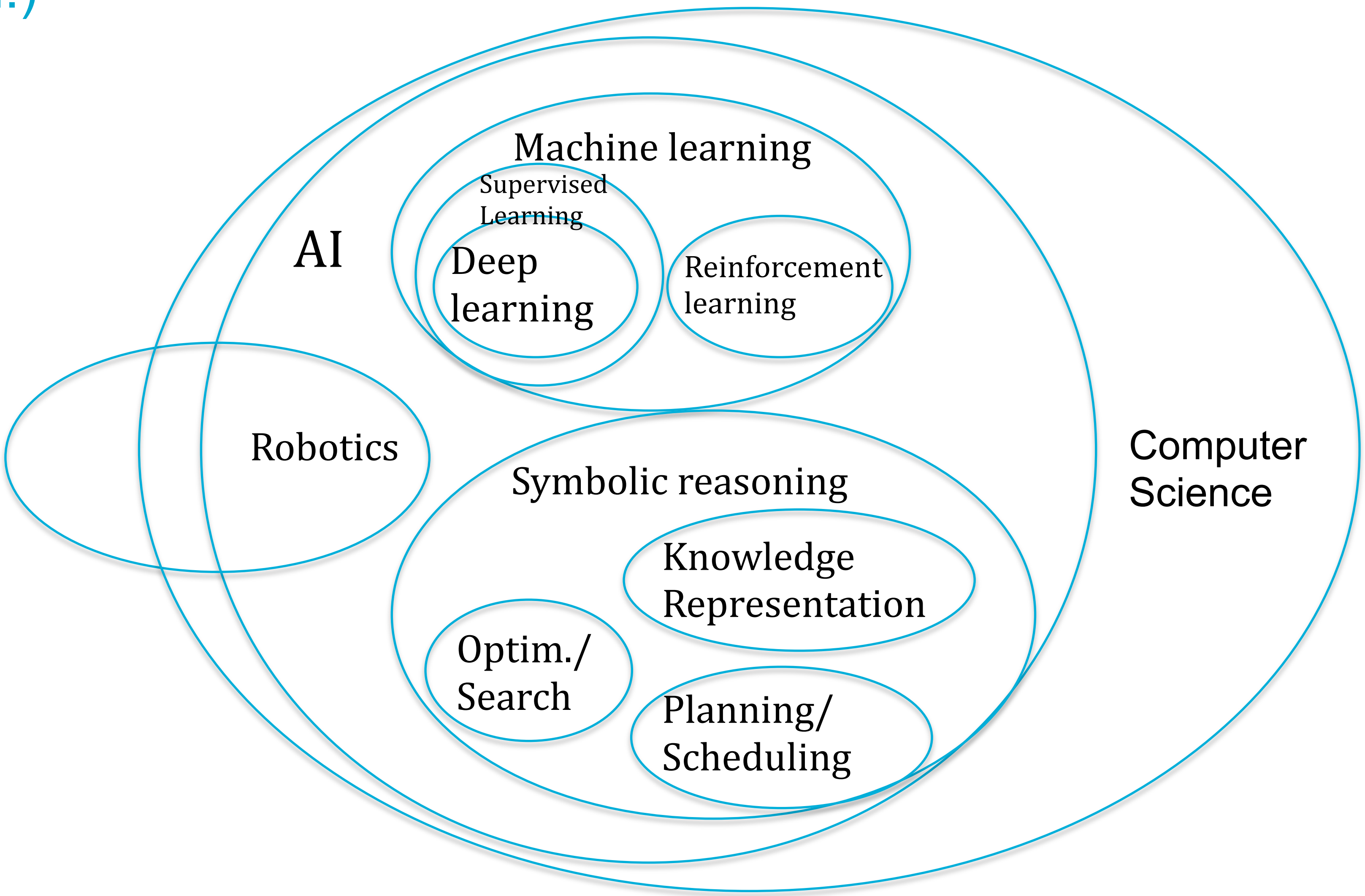


AI: Machine Learning

- Data driven
- Needs data curation
  - Unbiased, diverse, inclusive
- Agnostic algorithm whose parameters are set via training

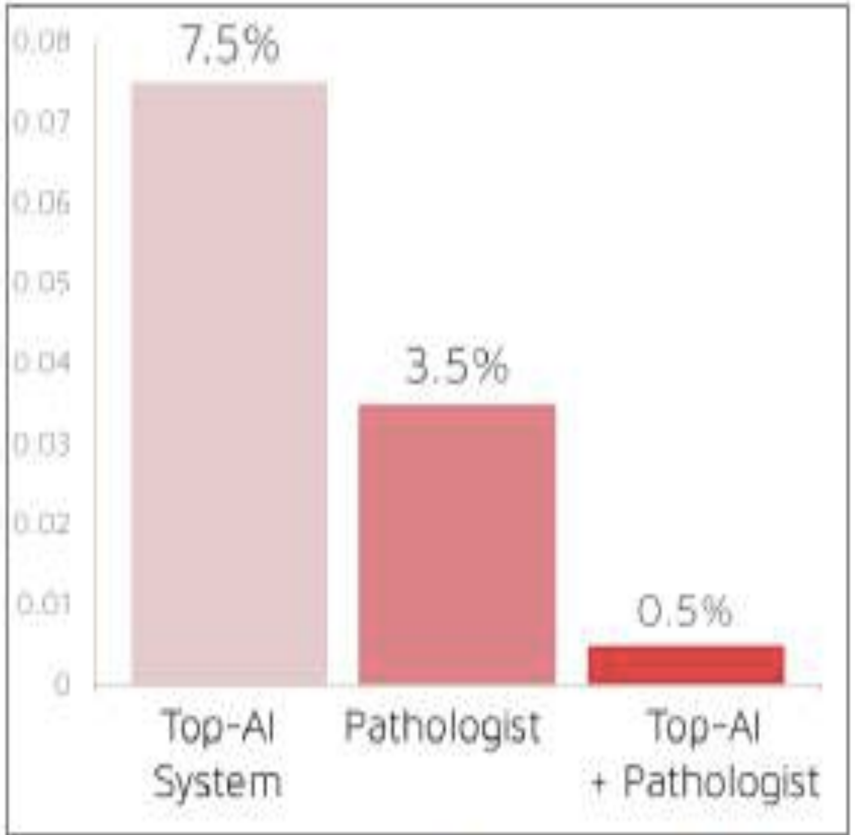
- Pros:
  - Flexible
  - Accurate also for ill-defined problems
- Cons:
  - Correlation rather than causation
  - Not always easy to provide “meaningful” explainations
  - Need huge amounts of data, and therefore computing power
  - Needs data curation
  - Adversarial attacks

AI and its subdisciplines
(very simplified!)



AI and people are very complementary

- We are better at
  - Asking questions and define problems to be solved
  - Common sense reasoning
  - Intuition
  - Creativity
  - Associations and analogies
- AI is better at
  - Handling huge amounts of data
  - Pattern discovery in data
  - Statistic and Probabilistic Reasoning



AI significantly reduces pathologist error rate in the identification of metastatic breast cancer from sentinel lymph node biopsies.

Enterprise AI

Developing and deploying AI to help other enterprises

- Needs to work with professionals
  - Helping people to do their job
- Heavily regulated domains
  - Healthcare, transportation, financial services, legal system
- A lot of domain knowledge
- Heavy use of natural language
  - Spoken and written
- Small amount of data
  - Solving new problems
- Human acceptance of the technology

AI actors – enterprise AI

Current limits of AI

- Common sense reasoning
- Combination of learnng and knowledge reasoning
- Natural language understanding
- Learning from few examples
- Learning general concepts

- Ethics-related limitations:
  - Bias → fairness
  - Black-box → explainability
  - Adversarial attacks→ robustness

Natural Language Understanding

- Winograd Schema challenge
  - Anaphora resolution

- «The box did not fit in the suitcase because it was too small/large»

- What is small/large?
  - Small → the suitcase
  - Large → the box

- The best AI systems have a 60% accuracy

Ethical issues in current AI: the age of trust

- Without trust there will not be full adoption, and therefore we will miss the huge positive effect of AI
  1. Trust in the AI technology
  2. Trust in those who produce AI
  3. Trust in those who regulate AI

- IBM IBV study on >1000 C-level executives and policy makers
  – 80% say that concerns about trust, privacy, and transparency are a barrier to AI adoption
  – 80% consider trusted training data important

Building trust on all dimensions

- Trust in AI
  - Bias in data or models: Is AI fair in its decisions?
  - Value alignment: Is AI understanding our intentions?
  - Explainability and transparency: How is it making decisions? How can I be sure of no deception or manipulation?
  - Robustness and safety/security: Can we make AI robust to adversarial examples and secure to attacks?
- Trust in AI producers
  - Data handling: How and for what purpose are my data used?
  - Design transparency: How can I assess the properties of the AI models I use?
- Trust in governments/policy makers
  - Personal data protection: Is my personal data going to be protected?
  - Privacy: Should we abandon online digital privacy to get better and better AI services?
  - Accountability: Who is to blame if something goes wrong?
  - Impact on jobs: How do we relocate and retrain people who lost their job to automation?
  - AI weaponization: Should AI be used to automate arms and fight against each other?

What is needed for trustworthy enterprise AI?

Transparent and explicit data policy

- Guidelines for developers
- Open-source initiatives
- AI ethics board/discussion/auditing mechanism

Technology properties (research/platforms/products):
- Accuracy
- Bias
- Value alignment

Explainability
- **Contextual and personalized**

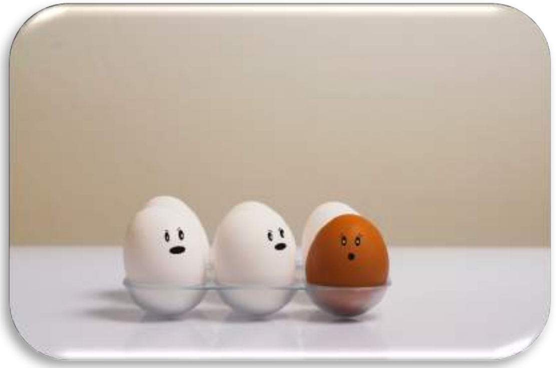
Design choices (AI factsheet)

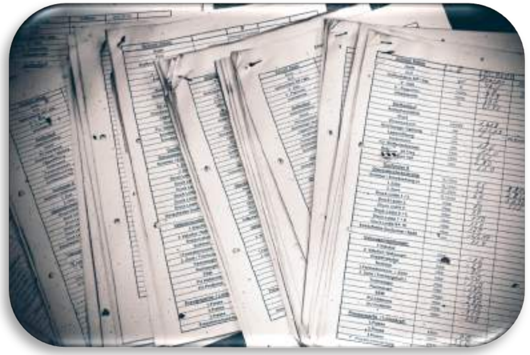
Education on using AI and embedding it in decision making process

Awareness, **education, community impact**

What does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)





Is it fair?
Is it aligned with my values?

Is it easy to understand?

Is it robust?

Is it accountable?

IBM's vision for Trusted AI

Pillars of trust, woven into the lifecycle of an AI application

FAIRNESS+

EXPLAINABILITY

ROBUSTNESS

ASSURANCE

supported by an instrumented platform

AI Lifecycle Manager

IBM's vision for Trusted AI

Pillars of trust, woven into the lifecycle of an AI application

FAIRNESS+

EXPLAINABILITY

ROBUSTNESS

ASSURANCE

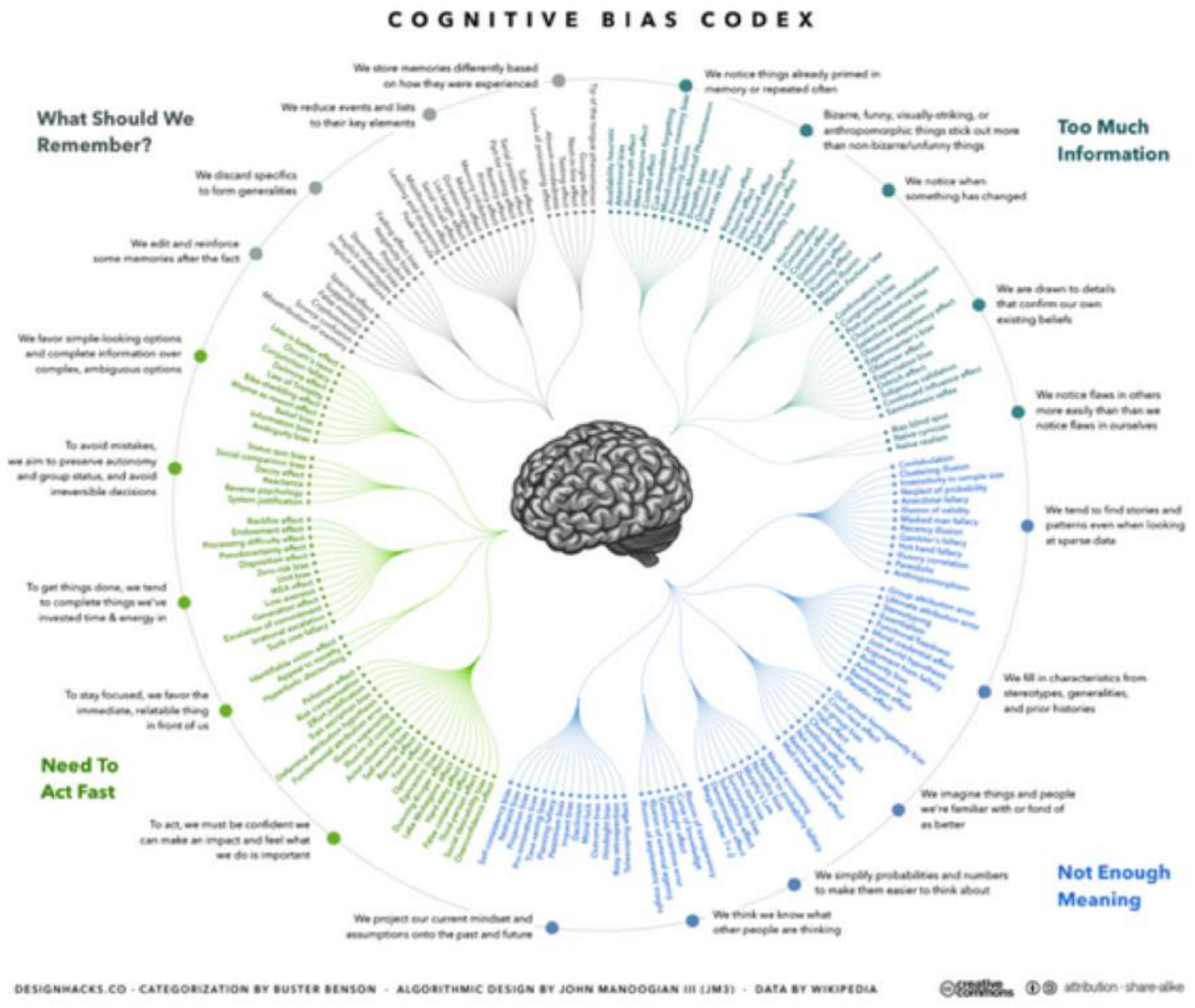
supported by an instrumented platform

AI Lifecycle Manager

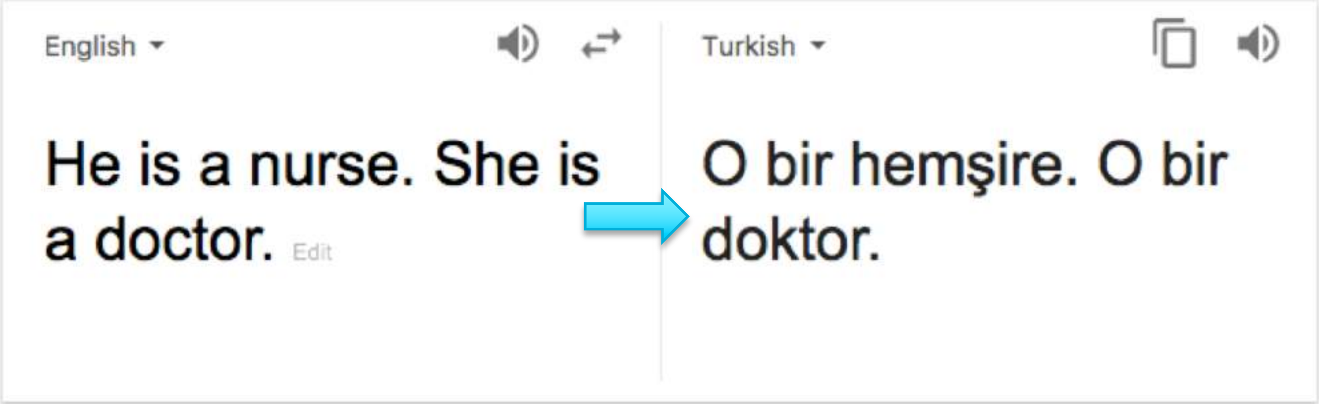
Bias in AI

- Bias: prejudice for or against something
- As a consequence of bias, one could behave unfairly to certain groups compared to others

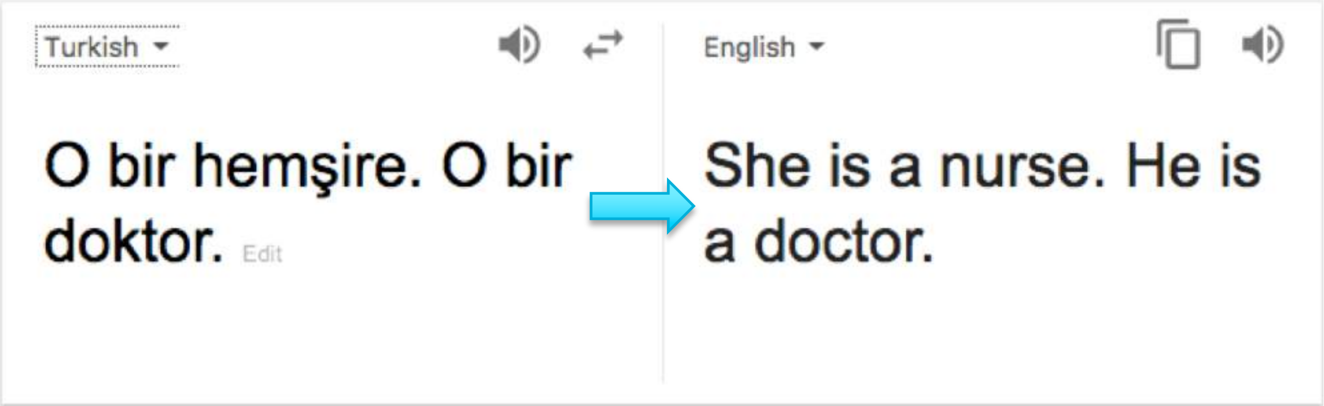
- Why should AI be biased?
  - Trained on data provided by people, and people are biased
  - Learning from examples and generalizing to situations never seen before



Language translation (2018)



English to Turkish



Turkish to English

AI Fairness 360

An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias

IBM Research Trusted AI

Home Demo Resources Community

AI Fairness 360 Open Source Toolkit

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing over 30 fairness metrics and 9 state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

API Docs ↗ Get Code ↗

Not sure what to do first? Start here!

Read More
Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.
→

Try a Web Demo
Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.
→

Watch a Video
Watch a video to learn more about AI Fairness 360.
→

Read a paper
Read a paper describing how we designed AI Fairness 360.
→

Use Tutorials
Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application domains.
→

Ask a Question
Join our AIF360 Slack Channel to ask questions, make comments and tell stories about how you use the toolkit.
→

View Notebooks
Open a directory of Jupyter Notebooks in GitHub that provide working examples of bias detection and mitigation in sample datasets. Then share your own notebooks!
→

Contribute
You can add new metrics and algorithms in GitHub. Share Jupyter notebooks show-casing how you have examined and mitigated bias in your machine learning application.
→

Learn how to put this toolkit to work for your application or industry problem. Try these tutorials.

Credit Scoring
See how to detect and mitigate age bias in predictions of credit-worthiness using the German Credit dataset.
→

Medical Expenditure
See how to detect and mitigate racial bias in a care management scenario using Medical Expenditure Panel Survey data.
→

Gender Bias in Face Images
See how to detect and mitigate bias in automatic gender classification of face images.
→

Web experience: http://aif360.mybluemix.net/
Code: https://github.com/IBM/AIF360
Paper: https://arxiv.org/abs/1810.01943

More than fairness: value alignment

- AI agents may misunderstand the real intention of the human
  - Lack of common sense knowledge
  - Data not inclusive or representative enough
  - Values non well defined or implicit

- This can bring AI agents to do unexpected and undesired actions

Examples of value misalignment

- An Eurisko game-playing agent that got more points by falsely inserting its name as the creator of high-value items
- A Lego staking system that flips the block instead of lifting, since lifting encouragement is implemented by rewarding the z-coordinate of the bottom face of the block
- A sorting program that always outputs an empty list, since it is considered a sorted list by the evaluation metric
- A game-playing agent that kills itself at the end of level 1 to avoid losing in level 2
- A robot hand that pretends to grasp an object by moving between the camera and the object
- A game-playing agent that pauses the game indefinitely to avoid losing


WIRED

TOM SIMONITE BUSINESS 08.08.18 09:00 AM

WHEN BOTS TEACH THEMSELVES TO CHEAT

List of 40+ examples: https://t.co/mAGUf3quFQ

Two explored solutions

- Recommendation systems
- Goal: to teach AI systems how to obey behavioral constraints learned by observation while still being responsive to the feedback from users
  - Reinforcement Learning approach
  - Examples to describe the ethical constraints, learnt offline
  - Constrained RL behavior during online use

- Preferences and ethical priorities
- Goal: To achieve personalization while not compromising essential values and principles
  - Preference frameworks (CP-nets) to model both preferences and ethical guidelines
  - Distance between CP-net structures
  - Distance thresholds to decide if agent can follow its preferences or must be better aligned to ethical priorities

Our vision for Trusted AI

Pillars of trust, woven into the lifecycle of an AI application

FAIRNESS+

EXPLAINABILITY

ROBUSTNESS

ASSURANCE

supported by an instrumented platform

AI Lifecycle Manager

But what is it that we are asking for?

The General Data Protection Regulation (GDPR)

- Limits to decision-making based solely on automated processing and profiling (Art.22)
- Right to be provided with meaningful information about the logic involved in the decision (Art.13 (2) f. and 15 (1) h)

Paul Nemitz, *Principal Advisor, European Commission*
Talk at IBM Research, Yorktown Heights, May, 4, 2018

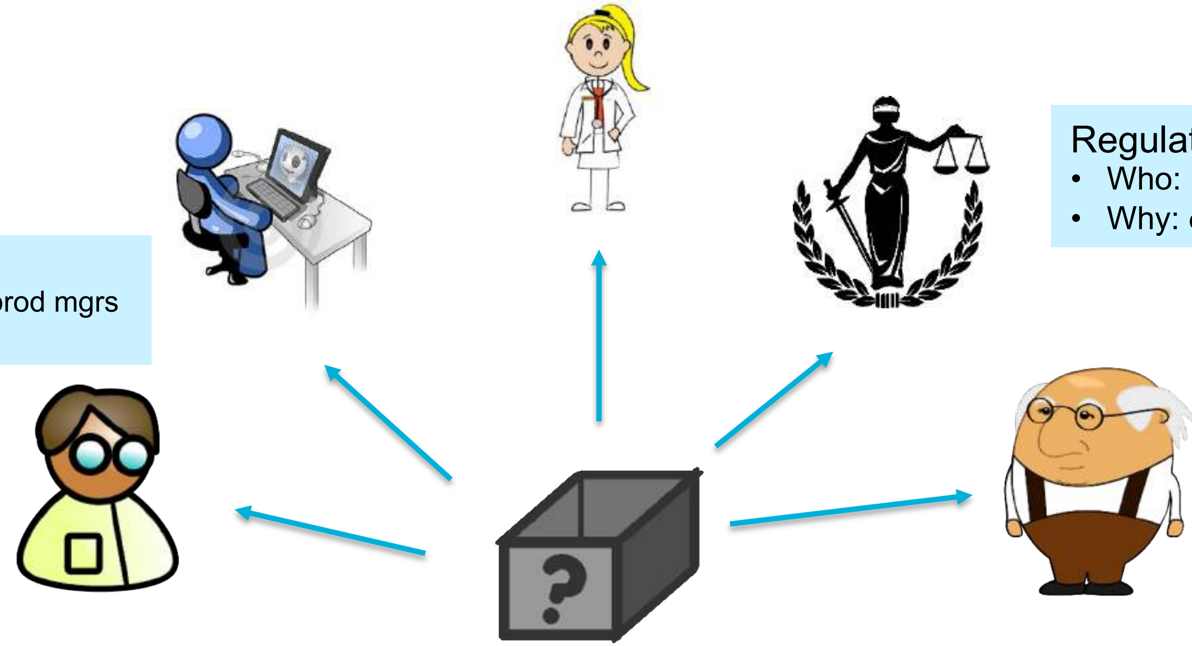
Meaningful Explanations Depend on the Explanation Consumer

End Users
- Who: Physicians, judges, loan officers, teacher evaluators
- Why: trust/confidence, insights(?)

Regulatory Bodies
- Who: EU (GDPR), NYC Council, US Gov't, etc
- Why: ensure fairness for constituents

AI System builders, stakeholders
- Who: data scientists, developers, prod mgrs
- Why: ensure/improve performance

Affected Users
- Who: Patients, accused, loan applicants, teachers
- Why: understanding of factors



Must match the complexity capability of the consumer
Must match the domain knowledge of the consumer

Three dimensions of explainability

One explanation does not fit all: There are many ways to explain things

directly interpretable vs. post hoc interpretation

directly interpretable

The oldest AI formats, such as decision rule sets, decision trees, and decision tables are simple enough for people to understand. Supervised learning of these models is directly interpretable.

post hoc interpretation

Start with a black box model and probe into it with a companion model to create interpretations. The black box model continues to provide the actual prediction while interpretation improve human interactions.

global (model-level) vs. local (instance-level)

global (model-level)

Show the entire predictive model to the user to help them understand it (e.g. a small decision tree, whether obtained directly or in a post hoc manner).

local (instance-level)

Only show the explanations associated with individual predictions (i.e. what was it about the features of this particular person that made her loan denied).

static vs. interactive (visual analytics)

static

The interpretation is simply presented to the user.

interactive (visual analytics)

The user can interact with interpretation.

Our vision for Trusted AI

Pillars of trust, woven into the lifecycle of an AI application

FAIRNESS+

EXPLAINABILITY

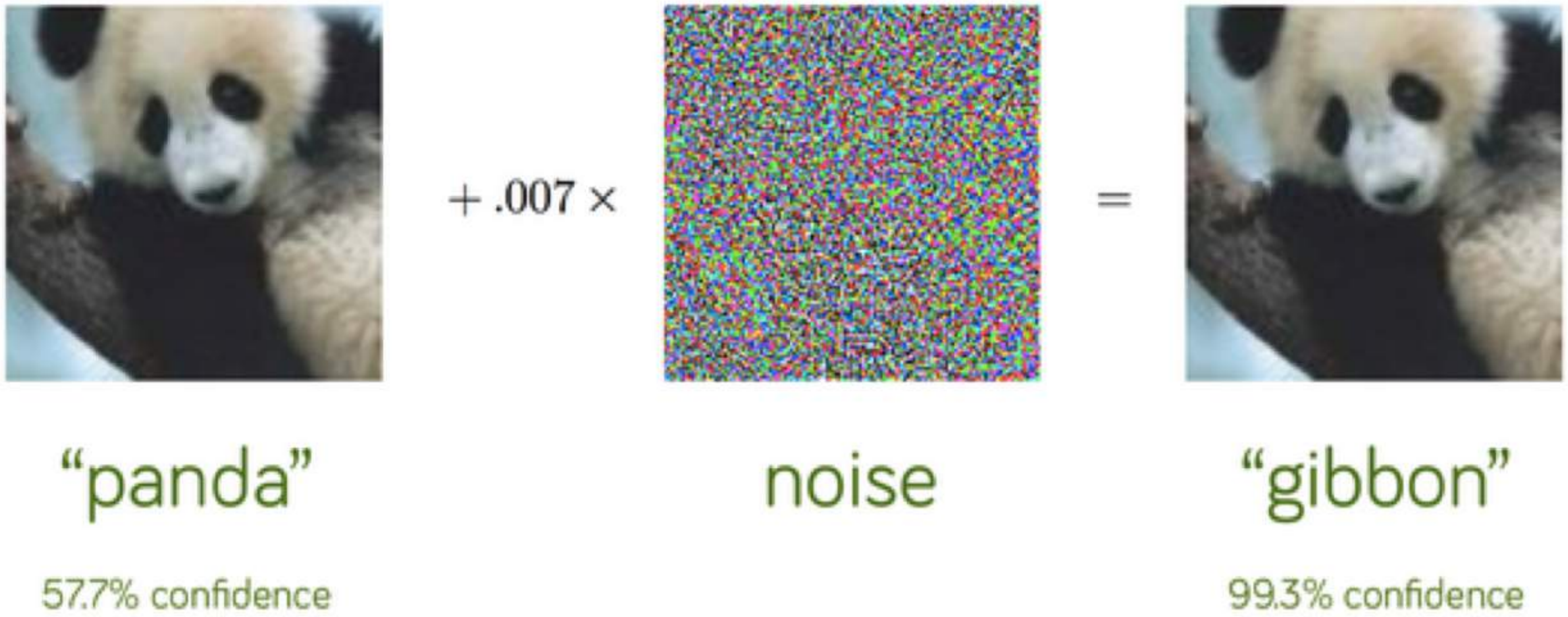
ROBUSTNESS

ASSURANCE

supported by an instrumented platform

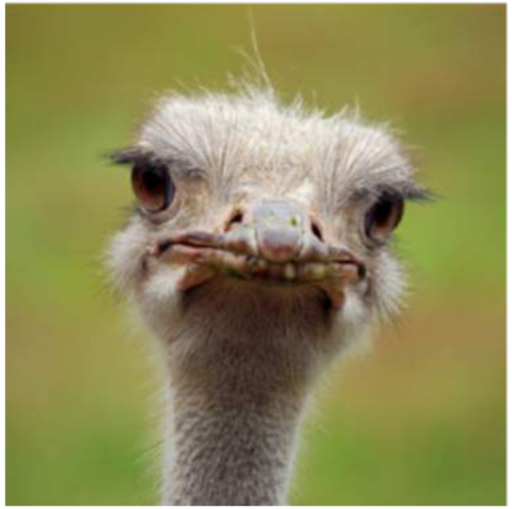
AI Lifecycle Manager

Adversarial Samples

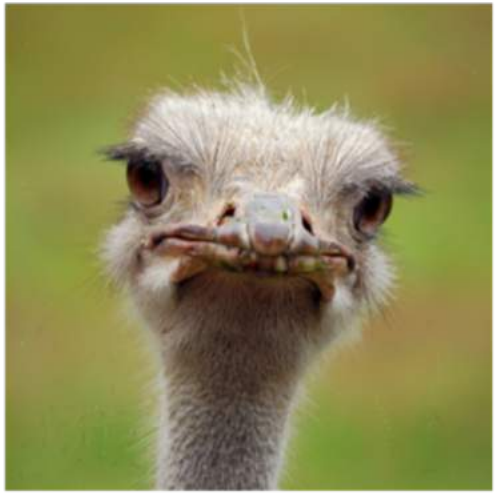


Adversarial Samples

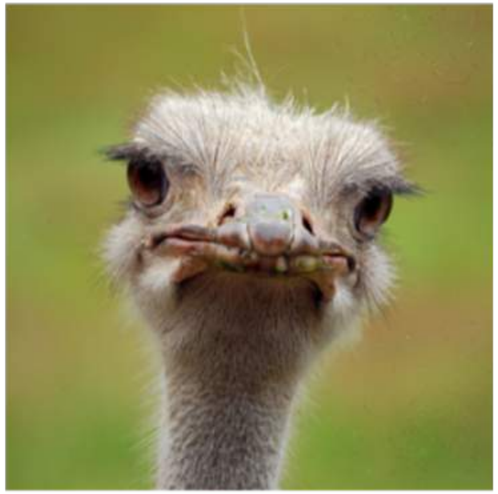
Ostrich



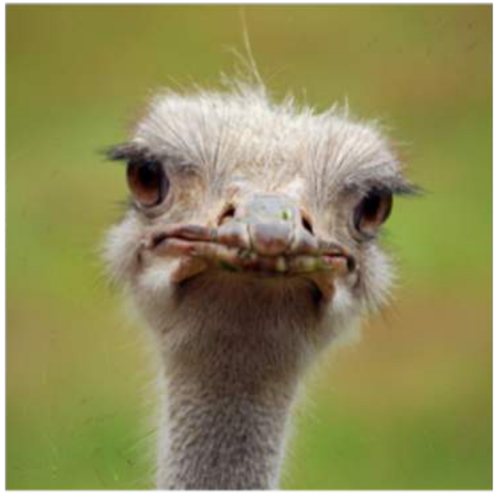
Safe



Shoe Shop



Vacuum

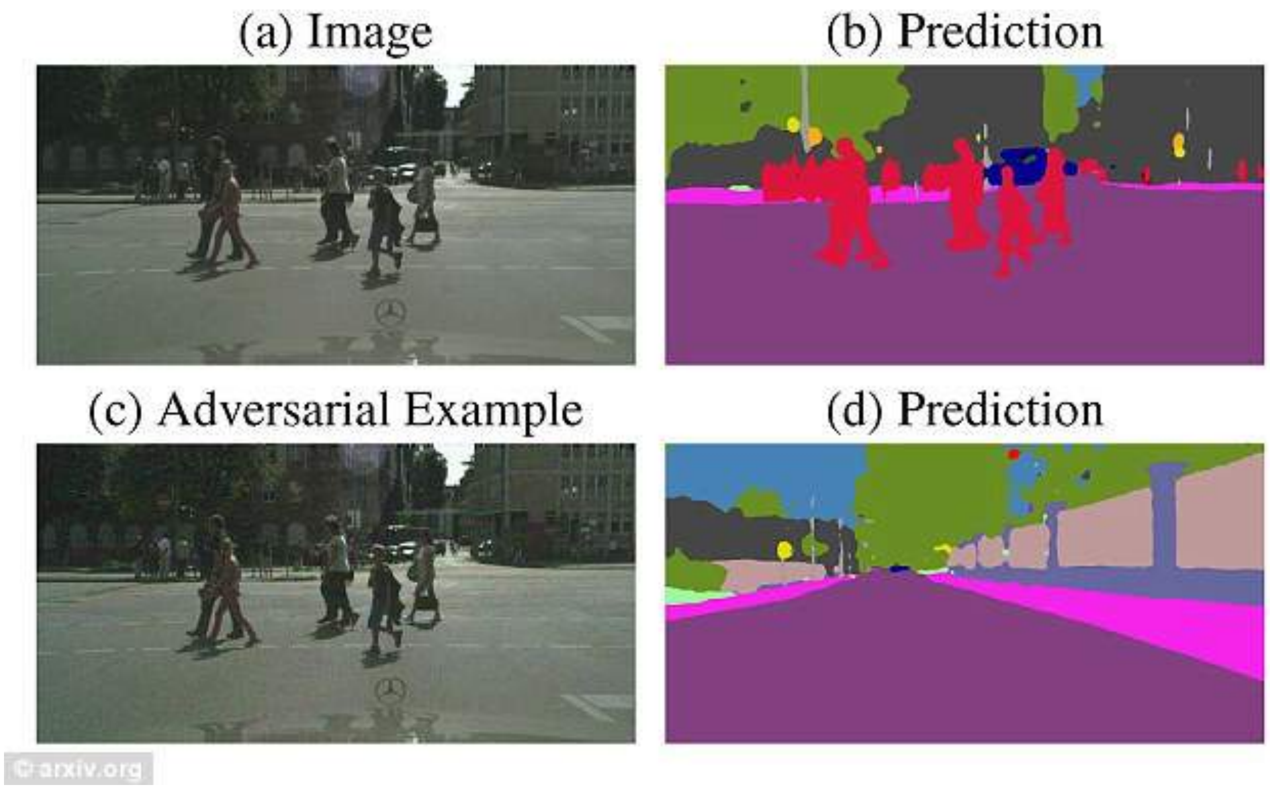


Attacks on AI

Poison training data and corrupt models

Evade detection by fooling models

Sharif et al., "Accessorize to a Crime - Real and Stealthy Attacks on State-of-the-Art Face Recognition," *CCS*, 2016.



[Jan Hendrik Metzen](), [Mummadi Chaithanya Kumar](), [Thomas Brox](), [Volker Fischer](). Universal Adversarial Perturbations Against Semantic Image Segmentation. arXiv 2017.

The Adversarial Robustness Toolbox

Adversarial Robustness
- Attack Agnostic Metrics
- Adversarial Sample Detection
- Input Preprocessing
- Model Hardening
- Robust Model Architectures

Model Theft
- Prevention of theft via APIs
- Detection of model theft attacks
- Deterring theft through model watermarking

Model and Data Privacy
- Provable privacy guarantees for training data (local differential privacy)
- Secure federated learning

Poisoning Attacks
- Detect poisoned training data and models
- Poison can degrade performance or insert backdoors
- Detection of poisoned samples at inference time

IBM **ART**
Adversarial Robustness Toolbox

Model Robustness Service
- Tooling layer to operationalize robustness in an easy-to-use service
- Supports defenses constructed from ART building blocks to evaluate robustness, harden vulnerable models, or repair poisoned models.
- Easily integrates into existing training/ModelOps pipelines (automated mode) and includes a GUI for exploration (interactive mode)

https://adversarial-robustness-toolbox.readthedocs.io/

Our vision for Trusted AI

Pillars of trust, woven into the lifecycle of an AI application

FAIRNESS+

EXPLAINABILITY

ROBUSTNESS

ASSURANCE

supported by an instrumented platform

AI Lifecycle Manager

Transparent reporting mechanism are basis for trust in many industries and applications

Nutrition Facts
Serving Size 8 oz
Servings Per Container 1.5

Amount Per Serving	
Calories 23	
	% Daily Value*
Total Fat 0g	**0%**
Saturated Fat 0g	**0%**
Trans Fat 0g	
Cholesterol 0mg	**0%**
Sodium 0mg	**0%**
Total Carbohydrate 5g	**2%**
Dietary Fiber 0g	**0%**
Sugars 6g	
Protein 1g	**2%**

*Percent Daily Values are based on a 2,000 calorie diet.

ENERGY STAR

Moody's		S&P		Fitch		Rating description	
Long-term	Short-term	Long-term	Short-term	Long-term	Short-term		
Aaa	P-1	AAA	A-1+	AAA	F1+	Prime	Investment-grade
Aa1		AA+		AA+		High grade	
Aa2		AA		AA			
Aa3		AA−		AA−			
A1		A+	A-1	A+	F1	Upper medium grade	
A2		A		A			
A3	P-2	A−	A-2	A−	F2		
Baa1		BBB+		BBB+		Lower medium grade	
Baa2	P-3	BBB	A-3	BBB	F3		
Baa3		BBB−		BBB−			
Ba1	Not prime	BB+	B	BB+	B	Non-investment grade speculative	Non-investment grade aka high-yield bonds aka junk bonds
Ba2		BB		BB			
Ba3		BB−		BB−			
B1		B+		B+		Highly speculative	
B2		B		B			
B3		B−		B−			
Caa1		CCC+	C	CCC	C	Substantial risks	
Caa2		CCC				Extremely speculative	
Caa3		CCC−				Default imminent with little prospect for recovery	
Ca		CC					
		C					
C		D	/	DDD	/	In default	
/				DD			
				D			

IEEE
IEEE STANDARDS ASSOCIATION

We have recently proposed ”factsheets” for AI services

FactSheets: Increasing Trust in AI Services
through Supplier's Declarations of Conformity

M. Arnold,[1] R. K. E. Bellamy,[1] M. Hind,[1] S. Houde,[1] S. Mehta,[2] A. Mojsilović,[1]
R. Nair,[1] K. Natesan Ramamurthy,[1] D. Reimer,[1] A. Olteanu,* D. Piorkowski,[1]
J. Tsay,[1] and K. R. Varshney[1]
IBM Research
[1]Yorktown Heights, New York, [2]Bengaluru, Karnataka

Abstract

Accuracy is an important concern for suppliers of artificial intelligence (AI) services, but considerations beyond accuracy, such as safety (which includes fairness and explainability), security, and provenance, are also critical elements to engender consumers' trust in a service. Many industries use transparent, standardized, but often not legally required documents called supplier's declarations of conformity (SDoCs) to describe the lineage of a product along with the safety and performance testing it has undergone. SDoCs may be considered multi-dimensional fact sheets that capture and quantify various aspects of the product and its development to make it worthy of consumers' trust. Inspired by this practice, we propose FactSheets to help increase trust in AI services. We envision such documents to contain purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers. We suggest a comprehensive set of declaration items tailored to AI and provide examples for two fictitious AI services in the appendix of the paper.

1 Introduction

Artificial intelligence (AI) services, such as those containing predictive models trained through machine learning, are increasingly key pieces of products and decision-making workflows. A service is a function or application accessed by a customer via a cloud infrastructure, typically by means of an application programming interface (API). For example, an AI service could take an audio waveform as input and return a transcript of what was spoken as output, with all complexity hidden from the user, all computation done in the cloud, and all models used to produce the output pre-trained by the supplier of the service. A second more complex example would provide an audio waveform translated into a different language as output. The second example illustrates that a service can be made up of many different models (speech recognition, language translation, possibly sentiment or tone analysis, and speech synthesis) and is thus a distinct concept from a single pre-trained machine learning model or library.

In many different application domains today, AI services are achieving impressive accuracy. In certain areas, high accuracy alone may be sufficient, but deployments of AI in high-stakes decisions, such as credit applications, judicial decisions, and medical recommendations, require greater trust in AI services. Although there is no scholarly consensus on the specific traits that imbue trustworthiness in people or algorithms [1, 2], fairness, explainability, general safety, security, and transparency are some of the issues that have raised public concern about trusting AI and threatened the further adoption of AI beyond low-stakes uses [3, 4]. Despite active research and development to address these issues, there is no mechanism yet for the creator of an AI service to communicate how they are addressed in a deployed version. This is a major impediment to broad AI adoption.

Toward transparency for developing trust, we propose a *FactSheet* for AI Services. A FactSheet will contain sections on all relevant attributes of an AI service, such as intended use, performance, safety, and security. Performance will include appropriate accuracy or risk measures along with timing information. Safety, discussed in [5, 3] as the minimiza-

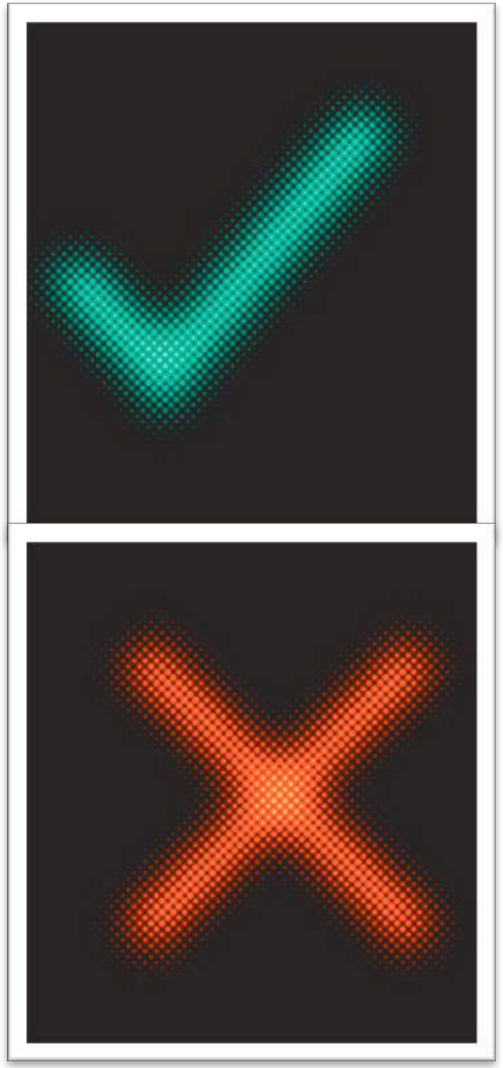
*A. Olteanu's work was done while at IBM Research. Author is currently affiliated with Microsoft Research.

1

- What is the **intended use** of the service output?
- What **algorithms** or techniques does this service implement?
- Which datasets was the service **tested** on?
- Describe the **testing methodology** and **test results**.
- Are you aware of possible examples of **bias, ethical** issues, or other **safety risks** as a result of using the service?
- Are the service outputs **explainable** and/or interpretable?
- For each dataset used by the service:
  - Was the dataset checked for **bias**?
  - What efforts were made to ensure that it is **fair** and **representative**?
  - Does the service implement and perform any **bias detection and remediation**?
- What is the **expected performance** on unseen data or data with different distributions?
- Was the service checked for **robustness against adversarial attacks**?
- When were the models last updated?

https://arxiv.org/abs/1808.07261

What information will be conveyed via a FactSheet?



The information reported on the FactSheet will depend on type of service, application domain, and user, but here are some examples:

- ❑ What is the intended use of the service output?
- ❑ What algorithms or techniques does the service implement?
- ❑ Which datasets was the service trained/tested on?
- ❑ Describe the testing methodology and results.
- ❑ How was the model trained, and were any steps taken to protect the privacy or confidentiality of the training data?
- ❑ Are you aware of possible examples of bias, ethical issues, or safety risks as a result of using the service?
- ❑ Does the service implement and perform any fairness checks detection and bias mitigation?
- ❑ What is the expected performance on data with different distributions?
- ❑ Was the service checked for robustness against adversarial attacks?
- ❑ When was the service last updated?
- ❑ Recommended uses. Not-recommended uses.

Our vision for Trusted AI

Pillars of trust, woven into the lifecycle of an AI application

FAIRNESS+

EXPLAINABILITY

ROBUSTNESS

ASSURANCE

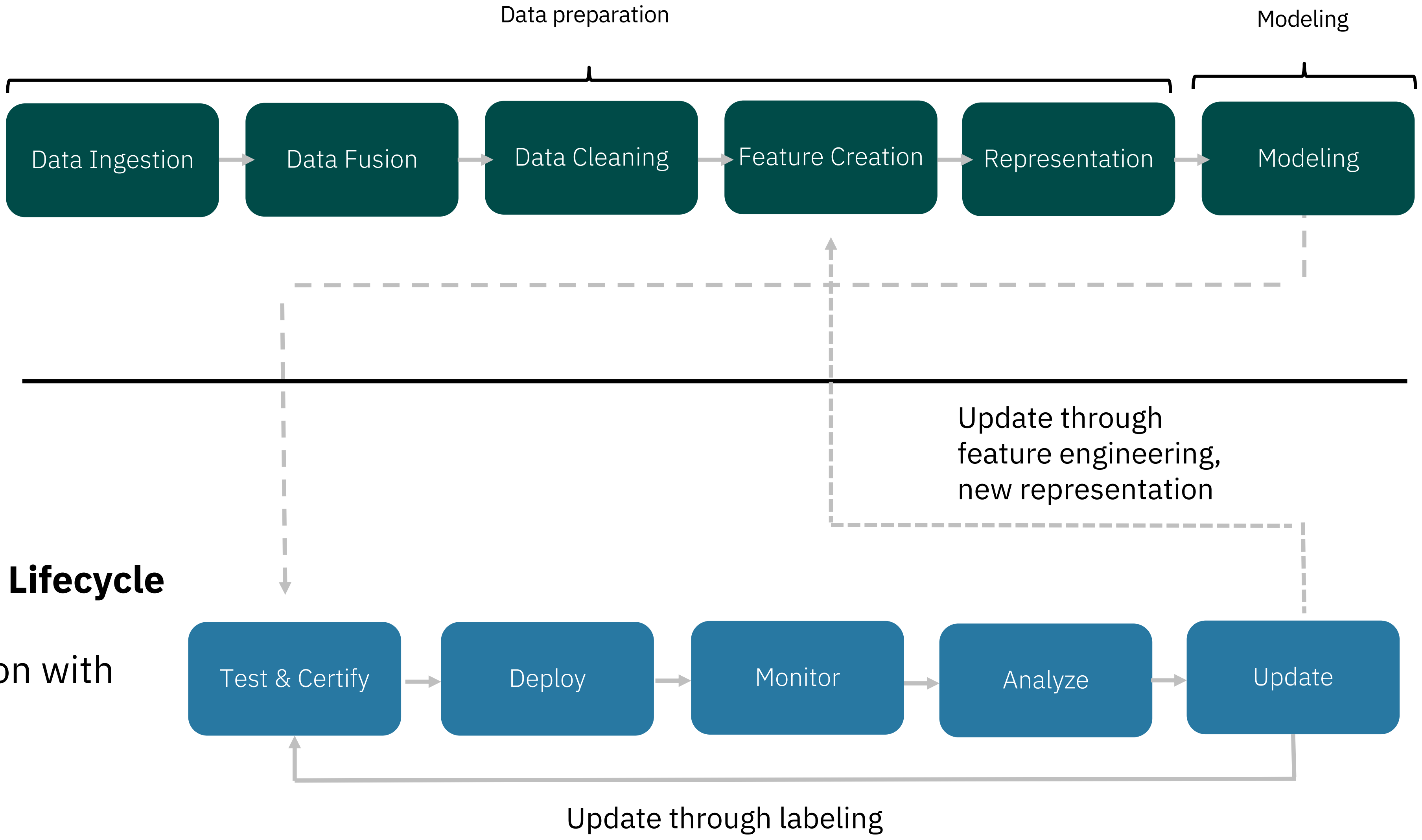
supported by an instrumented platform

AI Lifecycle Manager

What is the AI Lifecycle?

Data Science Lifecycle
- Data preparation and model creation

AI Deployment & Operations Lifecycle
- Model deployment and maintenance, in coordination with application



AI ethics at IBM: a holistic approach

Principles

Augmenting human intelligence, Trust and transparency (2017)

Policies

White paper on Learning to Trust AI systems (2016)
Data responsibility policy (2017)
Guidelines for designers and developers (2018)
European Union AI expert group membership (2018)
AI factsheet (2018)

Research and products

Bias detection, rating, and mitigation
Value alignment
Explainability
Robustness
AI fairness 360 toolkit (2018)
AI factsheet (2018)

Corporate responsibility

Impact of AI on jobs: skilling and reskilling, PTECH program

Internal coordination/awareness/driving/supporting mechanisms

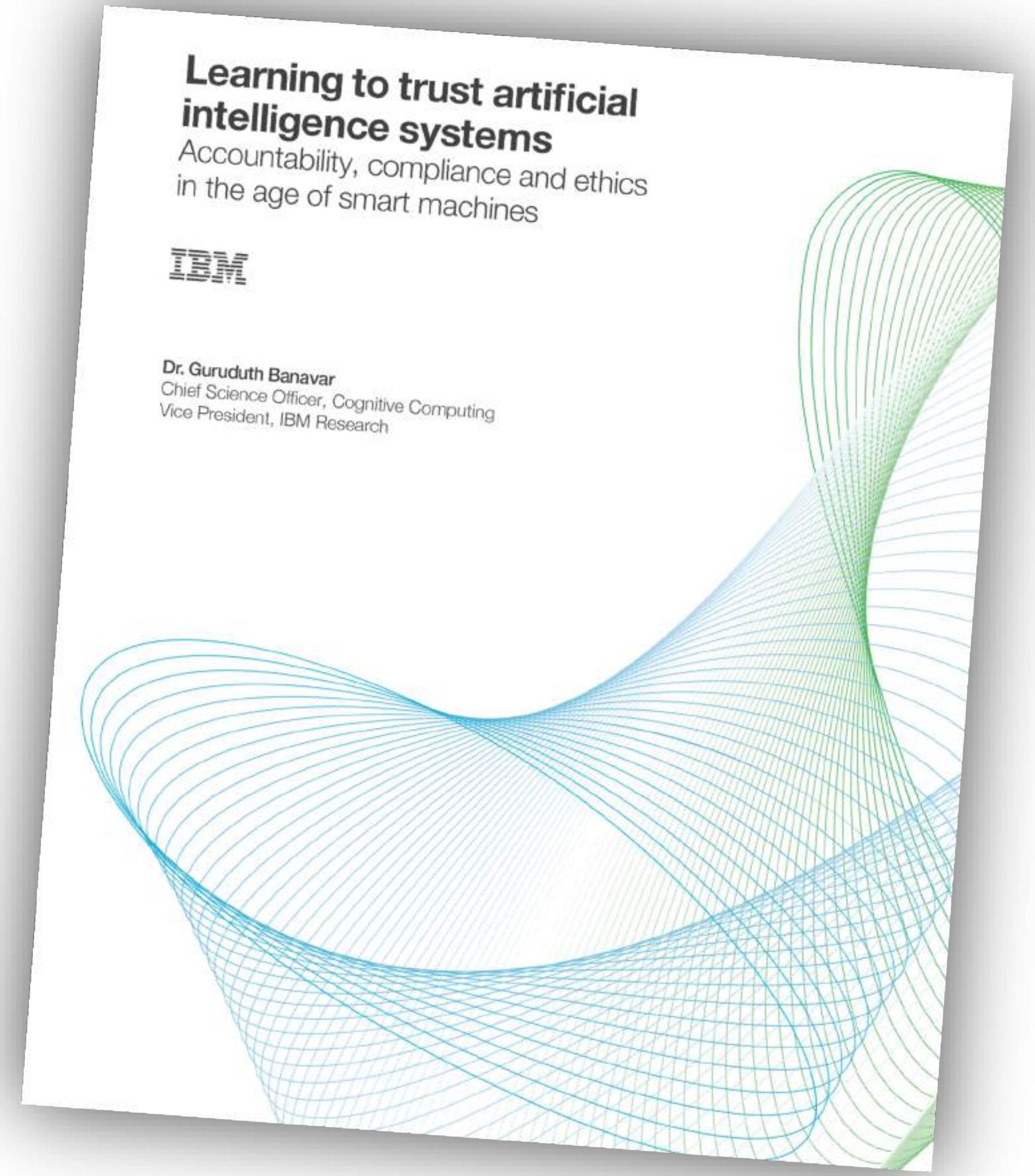
AI ethics coordination (research, products, policies, legal, communications)

Collaborative multi-disciplinary initiatives

IBM-MIT Watson AI Lab: theme on advancing shared prosperity with AI (2017- ongoing)
Founding partner of new AAAI/ACM conference on AI, Ethics, and Society (2018)
Founding partners of Partnership on AI (2016)
Executive committee membership of IEEE initiative on AI ethics (2017 - ongoing)
EU High Level Expert Group on AI (2018 – ongoing)
World Economic Forum partnership (ongoing)

White paper on AI ethics (2016)

1. Internal **IBM Cognitive Ethics Board**, to discuss, advise and guide the ethical development and deployment of AI
2. Company-wide **educational curriculum** on the ethical development of cognitive technologies.
3. IBM Cognitive Ethics and Society **research program** for the ongoing exploration of responsible development of AI systems aligned with our personal and professional values.
4. Participation in **cross-industry, government and scientific initiatives** and events around AI and ethics.
5. Regular, ongoing **engagements** with a robust ecosystem of academics, researchers, policymakers, NGOs and business leaders on the ethical implications of AI



Science for social good (since 2016)

https://www.research.ibm.com/science-for-social-good/

- AI and data science
- Summer fellowships for PhD students and postdocs
- Brings together
  - Research scientists and engineers
  - Academic fellows
  - Subject matter experts from a diverse range of NGOs
- To tackle emerging societal challenges using science and technology

Some examples:

- Opioid crisis
- Online hate speech
- Energy conservation
- Financial advisor for low-wage workers
- Illitteracy

Transparency and Trust in the Cognitive Era (Jan. 2017)

https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/

Purpose of AI

- To augment human intelligence
- Systems embedded in processes, systems, products and services by which business and society function
  - Should remain within human control

Transparency

- People need to have confidence in AI's recommendations, judgments and uses
- IBM will will make clear:
  - When and for what purposes AI is being applied
  - Major sources of data and expertise
  - Methods used to train those systems and solutions
  - Clients own their own business models and intellectual property
  - IBM will help clients to protect their data and insights

Skills

- IBM will work to help students, workers and citizens acquire skills and knowledge
  - To engage safely, securely and effectively in a relationship with AI cognitive systems
  - To perform the new kinds of work and jobs that will emerge in a cognitive economy

Data responsibility policy (Oct.2017)

https://www.ibm.com/blogs/policy/dataresponsibility-at-ibm/

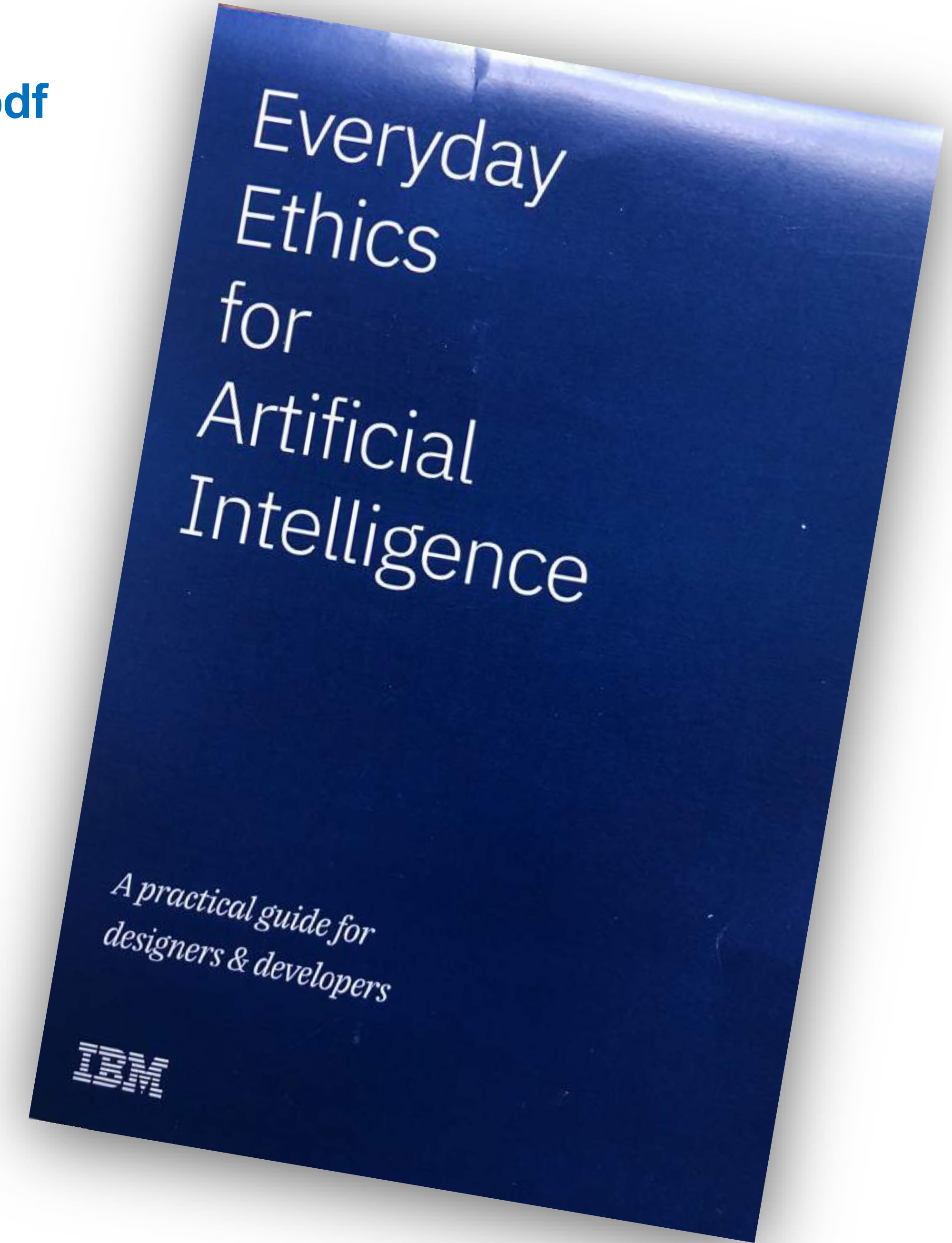
1. Data ownership and privacy
2. Data flows and access
3. Data security and trust
4. AI and data
5. Data skills and new collar jobs

Everyday ethics for AI – a practical guide for designers and developers (Sept. 2018)

https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

To help designers and developers think about AI ethics issues in their everyday work

- Accountability
- Value Alignment
- Explainability
- Fairness
- User Data Rights



Partnerships: a multi-stakeholder approach

AI producers

AI adopters

AI impacted users

Social scientists

Civil society


Partnership on AI
to benefit people and society
One organization
to develop and share the best practices for using and developing AI technologies and providing a global platform to discuss how AI will influence people and society.
7 Thematic Pillars
amazon
Microsoft
Google
DeepMind
IBM
Safety Critical AI
Fair, Transparent, and Accountable AI
AI, Labour and the Economy
Collaborations between People and AI systems
AI and Social Good
Social and Societal Influences of AI
Special Initiatives
Microsoft
IBM
DeepMind
Google
HARVARD Kennedy School
facebook
SAP
MacArthur Foundation
Cogitai
amazon.com
forum NEW ZEALAND
Berkeley UNIVERSITY OF CALIFORNIA
OpenAI
ACLU
AI2 ALLENAI.ORG
Data&Society
THE CENTRE FOR internet & society
cdt CENTER FOR DEMOCRACY & TECHNOLOGY
DIGITAL ASIA HUB
HUMAN RIGHTS WATCH
ebay
McKinsey&Company
CFI
intel experience what's inside
salesforce
Future of Humanity Institute
zalando
XPRIZE
unicef
Upturn
FUTURE OF PRIVACY FORUM
SONY
90+ Partners

IBM-MIT Watson AI Lab (since 2017)

http://mitibmwatsonailab.mit.edu/

- AI algorithms
- Physics of AI
- Applications of AI to industries
- Advancing shared prosperity through AI
    - AI ethics
    - AI for social good
    - AI and jobs

- Joint IBM-MIT projects

AAAI / ACM conference on

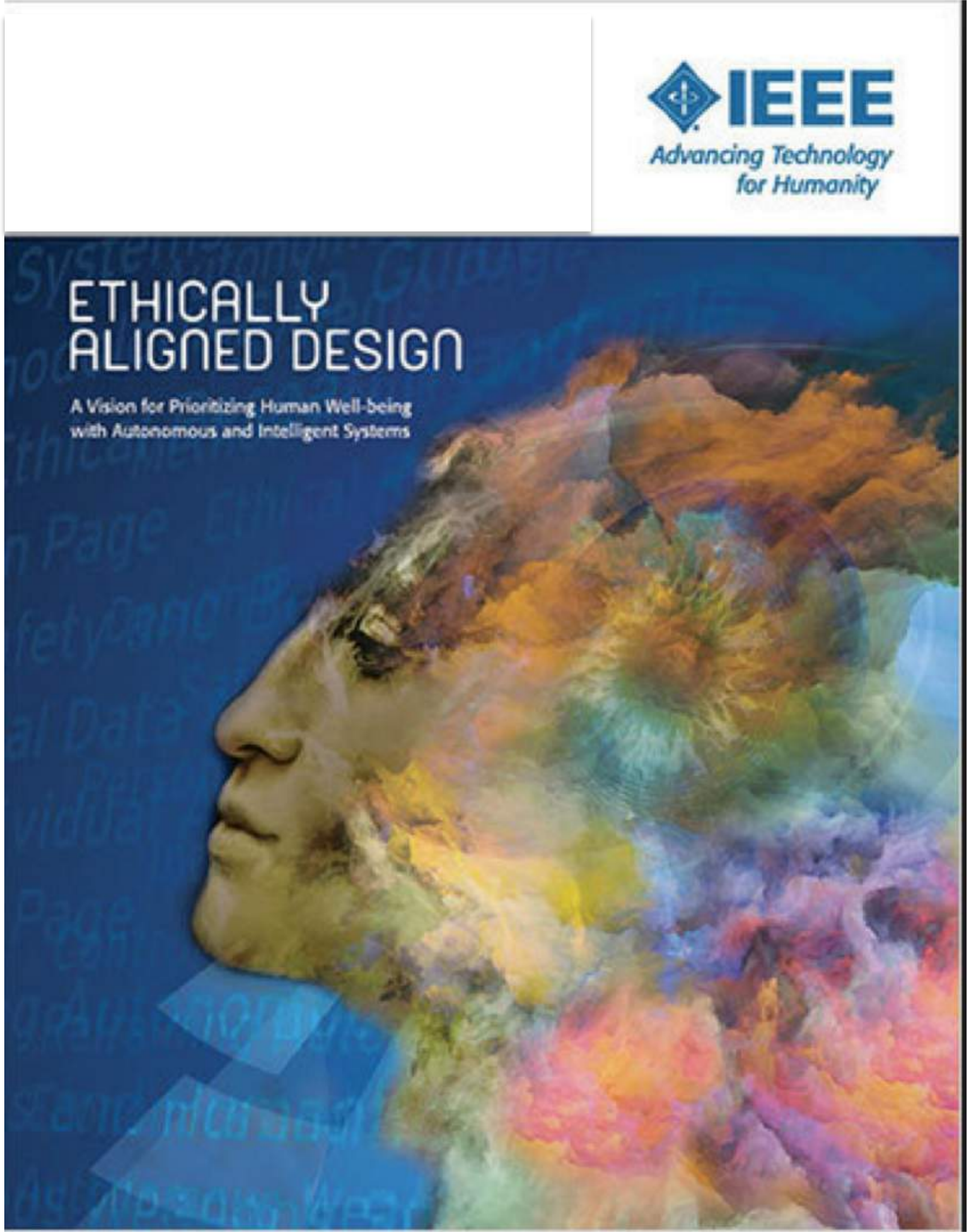
ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

- Multi-disciplinary scientific conference
- AAAI and ACM support
- Colocated with AAAI
- Started in 2018, second edition in Jan. 2018

IEEE Global Initiative on Ethics in Autonomous and Intelligent Systems (since 2016)

- About 250 global experts
- Feedback from anybody willing to comment
- About 300 pages
- Comprehensive and crowdsorced
- A chapter for each topic
  - List of issues and candidate recommendations on how to address them

Within the IEEE Standards Association
- Includes also the P700 series of standards
- Model Process for Addressing Ethical Concerns During System Design



EU Ethics Guidelines for AI

Human-centric approach

AI as a means, not an end

Trustworthy AI

foundational ambition

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

High-Level Expert Group and mandate

52 members from:

Industry

Academia

Civil society

Two deliverables

- Ethics Guidelines for Artificial Intelligence
- Policy & Investment Recommendations

Interaction with European AI Alliance

- Broad multi-stakeholder platform counting over 2800 members to discuss AI policy in Europe

Ethics Guidelines for AI – Intro

Trustworthy AI has three components

Lawful AI	Ethical AI	Robust AI

Three levels of abstraction

from principles (Chapter I)	to requirements (Chapter II)	to assessment list (Chapter III)

Ethics Guidelines for AI – Principles

4 Ethical Principles based on fundamental rights

Respect for human autonomy

Prevention of harm

Fairness

Explicability

Ethics Guidelines for AI – Requirements

- Human agency and oversight
- Technical Robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal & environmental well-being
- Accountability

To be continuously implemented & evaluated throughout AI system’s life cycle

Ethics Guidelines for AI – Assessment List

Assessment list to operationalise the requirements

- Practical questions for each requirement – 131 in total
- Test through piloting process to collect feedback from all stakeholders (public & private sector)

Official launch of piloting: 28 June – Stakeholder event

Moving forward with a holistic approach

- Technical innovation
  - From narrow capabilities to broader and deeper understanding
    - Focus on natural language
  - Value alignment
    - Including fairness and explainability
  - Combining learning ad reasoning
- Education
  - Tech students to consider the impact of what they will create
  - AI developers and operators
  - Policy makers on real AI capabilities, limitations, and issues
- Societal impact
  - Social scientists working with AI producers and policy makers
- Governance
  - Multi-stakeholder, multi-disciplinary, and multi-cultural discussion

Thanks!