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.

MIT Intro to Machine Learning course:
2013 – 138 students,
2016 – 302 students
2017 – 700 students
In 2018, blended AI will disrupt your customer service and sales strategy

75% of commercial enterprise apps will use AI by 2020

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

85% of CIOs will be piloting AI programs by 2020
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
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
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

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 s, from OPEN and save in CLOSED
4. Terminate: If s\in G then terminate with success and return f(s)
5. Expand: For each successor m of s
   - For each successor m, insert m in OPEN only if
     - if m \notin [OPEN\cup CLOSED]
       - set g(m)=g(n)+C(n,m)
       - Set f(m)=g(m)+h(m)
     - if m\in[OPEN\cup CLOSED]
       - set g(m)=\min(g(n), g(n)+C(n,m))
       - Set f(m)=g(m)+h(m)
   - If f(m) has decreased and m \notin 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

Forward Propagation
(Matrix Multiplication & Non-Linear Function)

Back Propagation
(Weight Update)

"cat"
"dog"
"horse"
"monkey"
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

![ImageNet Classification Error](chart)

- Deep Neural Networks
- GPU Hardware Accelerators

- 2010
- 2011
- 2012
- 2013
- 2014
- 2015
- Human
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” explanations
  - 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 learning 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
- Education on using AI and embedding it in decision making process
- Technology properties (research/platforms/products):
  - Accuracy
  - Bias
  - Value alignment
  - Explainability
  - Contextual and personalized Design choices (AI factsheet)
- Awareness, education, community impact
- Guidelines for developers
- Open-source initiatives
- AI ethics board/discussion/auditing mechanism
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

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
  - English: He is a nurse. She is a doctor.
  - Turkish: O bir hemşire. O bir doktor.

- Turkish to English
  - Turkish: O bir hemşire. O bir doktor.
  - English: She is a nurse. He is a doctor.
AI Fairness 360
An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias

Web experience: http://aif360.mybluemix.net/
Code: https://github.com/IBM/AIF360
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.

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

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

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

**AI System builders, stakeholders**
- **Who:** data scientists, developers, prod mgrs
- **Why:** ensure/improve performance
Three dimensions of explainability
One explanation does not fit all: There are many ways to explain things

**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.

**vs.**

**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)**

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).

**vs.**

**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**

The interpretation is simply presented to the user.

**vs.**

**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

Supported by an instrumented platform
AI Lifecycle Manager
Adversarial Samples

“panda”  + .007 × noise = “gibbon”

57.7% confidence  99.3% confidence
Adversarial Samples

Ostrich  Safe  Shoe Shop  Vacuum
Attacks on AI

Poison training data and corrupt models

Evade detection by fooling models

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

**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

supported by an instrumented platform
AI Lifecycle Manager
Transparent reporting mechanism are basis for trust in many industries and applications
We have recently proposed "factsheets" for AI services

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
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

Update through labeling

Update through feature engineering, new representation
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 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)
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)


- 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
- Illiteracy
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 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)

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
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

✓ Safety
Critical AI

✓ Fair, Transparent,
and Accountable AI

✓ AI, Labour and the
Economy

✓ Collaborations
between People
and AI systems

7 Thematic Pillars

AI and Social Good

Social and Societal
Influences of AI

Special Initiatives

90+ Partners

© 2018 International Business Machines Corporation
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
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 crowdsourced
- 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

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 and 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!