AI ethics

Francesca Rossi
IBM Research

IBM AI Ethics Global Leader
AAAI President-elect
ARTIFICIAL INTELLIGENCE
Intelligent algorithms defined and coded by people into machines

MACHINE LEARNING
Ability to learn without being explicitly programmed

DEEP LEARNING
Learning based on Deep Neural Networks

AI Day, Nov. 2020
## The Future Of/With AI

<table>
<thead>
<tr>
<th>General</th>
<th>• Able to solve many different problems and adapt to new environments</th>
</tr>
</thead>
</table>
| Trustworthy | • Reliable, robust, explainable, aware of its bias, compliant with ethics principles  
• Also in presence of uncertainty and disruptive events |
| Collaborative | • Fosters effective human-machine teaming  
• Especially in complex scenarios with high-stake decisions  
• Cooperation and autonomy |
| Sustainable and computationally scalable | • Energy and data efficient  
• Real-time decisions |

AI Day, Nov. 2020
Current AI: capabilities, limitations, ethical issues

### Capabilities

#### Machine Learning
- Learning from data (Deep, Reinforced, Supervised/Unsupervised/Self Supervised)
- Hidden patterns in huge amounts of data
- Prediction, perception tasks
- Correlation, pattern discovery, data mining
- Flexible, can handle uncertainty

#### Rule-based, symbolic, and logical approaches
- Explicit procedure to solve a problem
- Reasoning, planning, scheduling, optimization for complex problems
- Symbolic, traceable, explainable

### Limitations

- Generalizability and Abstraction
- Robustness and Resiliency
- Contextual awareness
- Multi-agent cooperation
- Resource efficiency (examples, energy, computing power)
- Adaptability
- Causality

### AI ethics issues

- Trust
- Fairness, robustness, explainability, causality, transparency
- Data governance, privacy, liability, human agency, impact on work and society
- AI autonomy vs augmented intelligence
- Real vs online life, metrics of success/goals
AI research: Europe vs USA vs Asia

- Europe – rules, less funds
  - Traditionally more on theoretical/rigorous/symbolic side
  - Applications mostly in the B2B space
- USA – funds, innovation
  - More applied/experimental/data driven side of AI
  - B2C and B2B business models
- Asia – data, funds
  - Mostly applied
  - More and more research, mostly data-driven
  - B2C
AI ethics

Multi-disciplinary and multi-stakeholder field of study

- Technical solutions: algorithms, toolkits, libraries
- Non-technical solutions: Guidelines, best practices, standards, incentives, funds, audit, certifications, policies, laws, governance, education

- AI needs data
  - Data privacy and governance
- AI is often a black box
  - Explainability and transparency
- AI can make autonomous decisions
  - Fairness and value alignment
- AI is based on statistics and has always a small percentage of error
  - Accountability
- AI can infer our preferences and manipulate them
  - Human and moral agency
- AI is very pervasive
  - Larger negative impacts for tech misuse
  - Impact on jobs and society

AI Day, Nov. 2020
AI PRINCIPLES in the world

Actors:
- Private sector
- Inter-governmental
- Multistakeholder
- Governments
- Civil society

Main themes:
- Human rights
- Human values
- Responsibility
- Human control
- Fairness
- Transparency and explainability
- Safety and Security
- Accountability
- Privacy

Principled AI Project, Berkman Klein’s Cyberlaw Clinic, 2019
Research venues focused on AI ethics
From AI ethics principles to practice: a multi-dimensional path
Global multi-stakeholder initiatives around AI and AI ethics:

Given the current state of AI, how can we address limitations and ethics issues?
Companies, universities, experts, associations, professional organizations

Partnership on AI, 2017-
- Geography: Asia 7%, Canada 3%, Europe 12%, Global 8%, UK 3%, USA 63%, Other 3%
- Sharing and defining best practices, lessons learnt, use-cases
- 100 partners: 20 companies, plus NGOs, civil society orgs, UN agencies, Univ., research centers
- “Productive disagreement”

High Level Expert Group on AI, 2018-2020
- Independent, multi-stakeholder
- Trustworthy AI, ethical guidelines
- Self assessment list
  - IBM AI factsheet
  - General and sector-specific recommendations for policy and investments

IEEE global initiative on Ethics of AIS, 2017-
- Ethically Aligned Design, 200+ pages, 200+ experts
- P7000 standard working groups
- AI certification

AI Day, Nov. 2020
Alternative approach to global collaboration:

Given a vision of the future, what is the correct trajectory for AI? UN agencies/countries, based on UN SDGs

<table>
<thead>
<tr>
<th>AI for Good Global Summit, 2017-</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Problem owners (UN agencies) and problem solvers</td>
</tr>
<tr>
<td>• 17 UN SDGs</td>
</tr>
<tr>
<td>• Global challenges (one just launched on gender equity, SDG5)</td>
</tr>
<tr>
<td>• XPRIZE partnership</td>
</tr>
<tr>
<td>• Another example: IBM Call for Code</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Global Partnership on AI (GPAI), 2020-</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Responsible development of AI</td>
</tr>
<tr>
<td>• Human rights, inclusion and diversity</td>
</tr>
<tr>
<td>• Innovation and economic growth</td>
</tr>
<tr>
<td>• Countries</td>
</tr>
<tr>
<td>• Started by France and Canada</td>
</tr>
<tr>
<td>• Also Australia, EU, Germany, India, Italy, Japan, Mexico, New Zealand, Korea, Singapore, Slovenia, UK, USA</td>
</tr>
</tbody>
</table>
Regulating AI

- Risk-based approach
- Focus on uses, not the technology
- Definition of high-risk
  - Sector-based or more fine-grained
- EU: future regulation only for high-risk uses of AI
- Self-assessment, standards, labelling for uses with less risk
Thanks!