Human-AI Collaboration: Technical & Ethical Challenges

FRANCESCA ROSSI
AI Ethics Global Leader
Distinguished Research Staff Member
IBM Research AI
Professor of Computer Science
University of Padova
Artificial Intelligence: Augmenting Human Intelligence

AI includes:
• Learning
• Reasoning
• Perceiving the environment
• Interacting with humans

Algorithms + Big data + Computing power

AI is already here:
• Web search
• Language translation
• Digital assistants
• Image understanding
• Text understanding and generation

Impact on all sectors:
• healthcare
• transportation,
• manufacturing,
• education
• ...

...
AI: augmenting human intelligence
AI: augmenting human intelligence

Human + Machine

Self-directed goals  Large-scale math
Common sense      Pattern discovery
Value judgment     Statistical reasoning

... leads to better decision making: more efficient, more confident, more informative, more grounded, more insightful, less resource-intensive, more ethical, less biased
Human + Machine: AI Challenges

Effective teamwork and natural interaction
• Dialog, deep NL understanding
• Common sense reasoning
• Theory of mind, affective computing, sentiment analysis

Proactive decision support
• Understanding what humans need, when, and in which form

Dynamic improvement over time
• Learning from few examples, learning concepts, transfer learning
Human + Machine: Ethical Challenges

Value alignment
• Modelling ethical principles and embedding them into AI systems’ machinery

Trust/Explanations/Interpretability
• Asking the “why” question and getting meaningful explanations/evidence
• Every AI system should be equipped with explanation capabilities

Bias/fairness
• Biased training data may lead to biased decisions
• Recognizing bias and helping people avoiding it
Human + Machine: helping humans

• No universal set of values
• Human behavior not always ethical nor rational

Proactive behavior:
• Alerting when humans deviate from ethical standards or introduce bias
• Measuring deviation
• Suggesting more ethical actions or mitigate bias
IBM’s Principles for the Cognitive Era

PURPOSE
To augment, not replace, human intelligence.

TRUST & TRANSPARENCY
In the development and deployment of AI systems, and the handling of data.

SKILLS
Responsibility to education to support workforce evolution.
Multidisciplinary Initiatives toward Beneficial AI