

G7 UCT and Industry Multistakeholder Conference

Trustworthy and Beneficial AI

FRANCESCA ROSSI

Distinguished Research Staff Member
Ethics of AI Global Leader
IBM Research

Professor of Computer Science
University of Padova



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

 @frossi_t

AI: augmenting human intelligence

Human + Machine

Self-directed goals

Large-scale math

Common sense

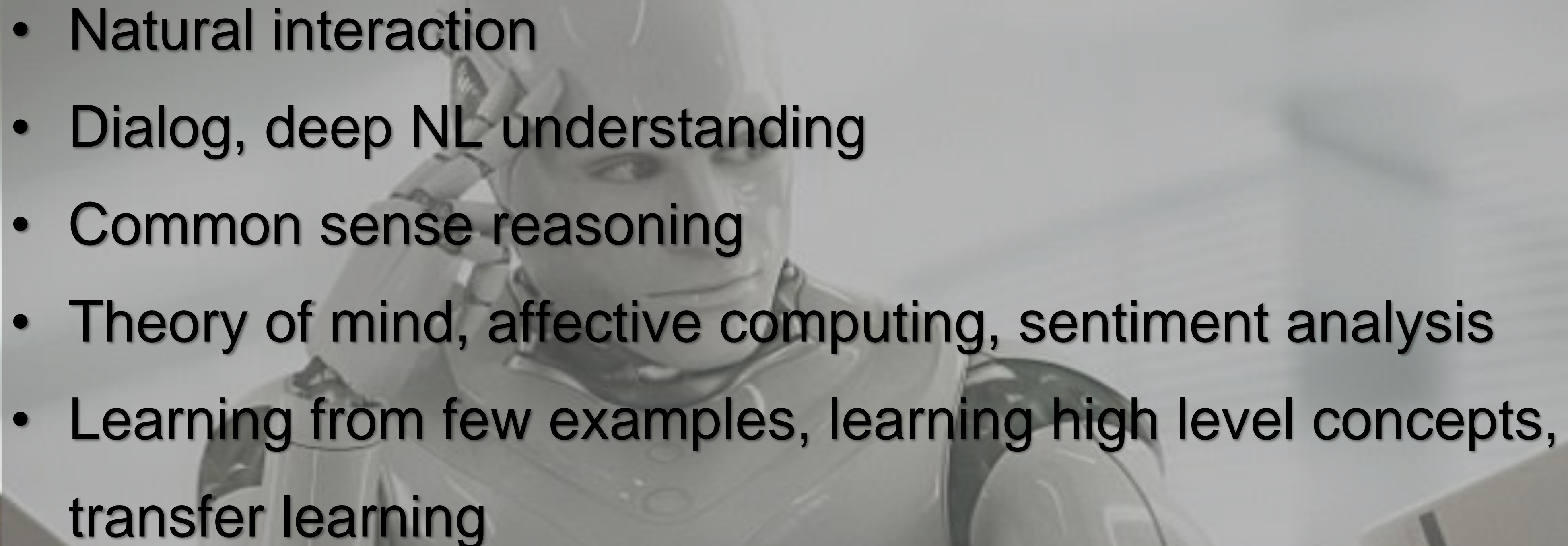
Pattern discovery

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

- Natural interaction
 - Dialog, deep NL understanding
 - Common sense reasoning
 - Theory of mind, affective computing, sentiment analysis
 - Learning from few examples, learning high level concepts, transfer learning
- 

Human + Machine: Ethical challenges

Value alignment

- Is it following my same ethical principles?

Trust

- Can I trust the decisions it makes/suggests me to make?

Explanations

- Why is it making/suggesting a certain decision?

Transparency

- Can I check its reasoning process?
-



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



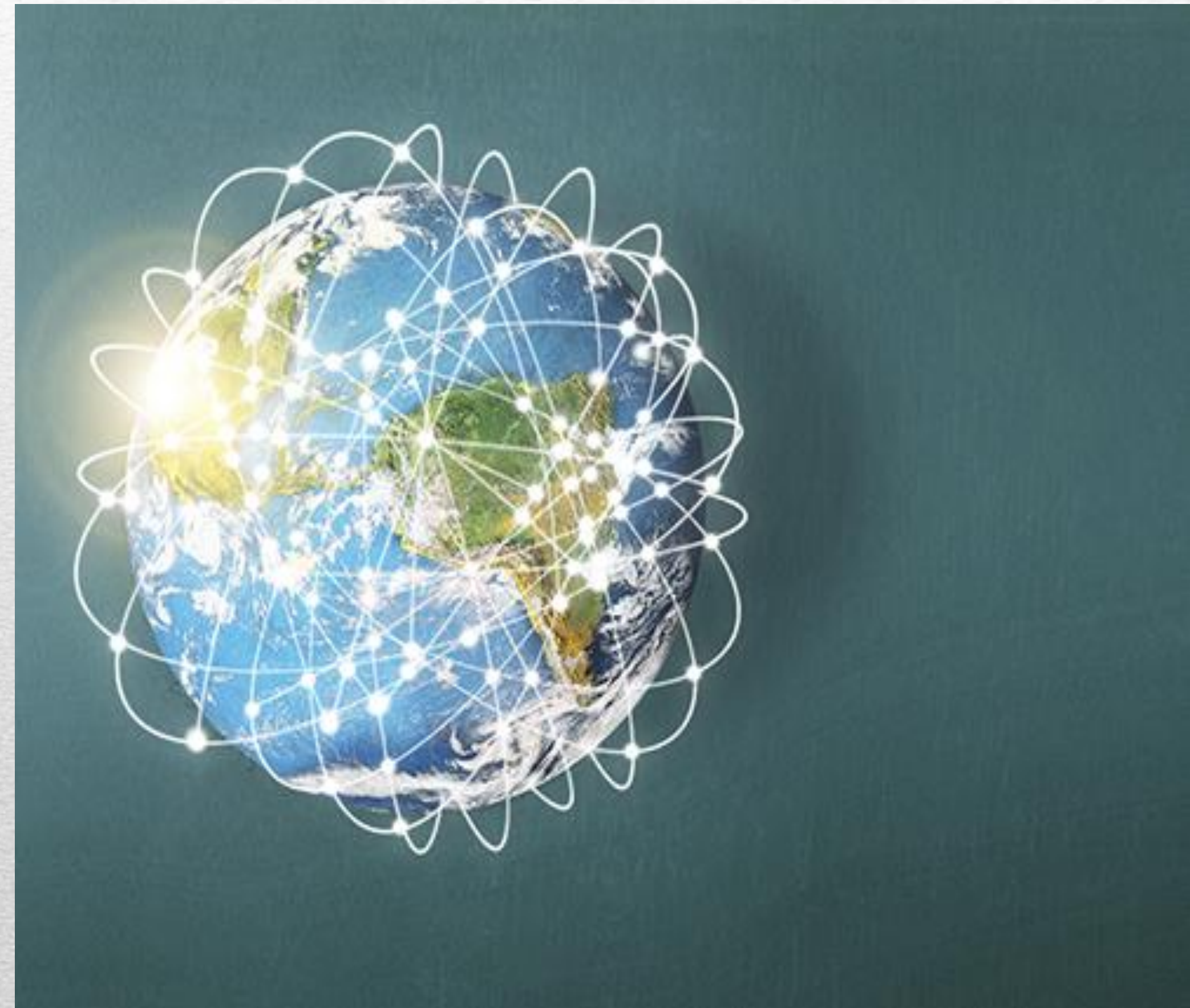
Ethical issues around the use and development of AI

Impact on

- Workforce
- Society
- Human-human interaction
- Education

Trusting AI producers

- Transparency in handling clients' data, privacy, ownership, storage, bias
- Ex.: IBM's principles for the cognitive era



Multidisciplinary Initiatives toward Beneficial AI



**K&L Gates Endowment for
Ethics and Computational Technologies**



UC Berkeley
Center for Human-Compatible AI



**USC Center for
Artificial Intelligence in 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 Pillar



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



30+ Partners

