Trust, Trustworthy and Autonomy

Zhe Liu (Peking University)

Geneva, Ma., 2018

Relevant Technological Development in AI & Robotics

1) Machine learning with big data

2) Biomechatronics

3) The increase of automation and 'autonomy'

4) Closer interaction between humans and AI & Robots (co-bots, cyber-crews, digital twins, cyborgs and so on)

Problems of Trust in AI

1) Mistrust & Overtrust: Interactions between humans and AI or robots

2) Trustworthy of AI & Robots
Interrupted communication between humanities people and AI scientists

Distrusted overseeing of designers, manufacturers and users

3) Open question of the relation between trust and trustworthy of AI & Robots

Deception in the interaction between Humans and AI/Robots

1) Personal-relationship-like interactions of humans with AI or robots

2) Artificial care and artificial friendship

3) Overtrust and mistrust of AI

4) Deception in the form of a genuine counterfeited companionship

Challenge to both individual Autonomy and Priniciplled Autonomy

Onora O'neill's valuable differentiation
1) Individual autonomy: independence of action in the causal chain of natural states and events (or desires and beliefs)

2) Priniciplled autonomy: not in any distinctive form of personal independence or self-expression but in the Kantian sense of non-derivative legislation or self-legislation
we must act on principles others can follow

3) Can the danger of Manipulation/Nudging/Deception by AI or Robots be mitigated on the basis of modern morality of autnomy in whatever form?

Possible loss of tustworthy

1) Humanities people do not have sufficient knowledge of what is going on in AI & robotics and hence raise unreasonable suspects of trustworthy of AI

2) The traditional dichonomy of fact and value

3) Tustworthy could be constructed on the basis of satisfaction of individual desires by AI

Paternalism v.s. Democracy

1) The traditional notion of risk

2) Paternalist mindset prevalent among many Chinese AI scientists and roboticists

3) Overseeing will, if not destroy, at least delay the techonological development in AI

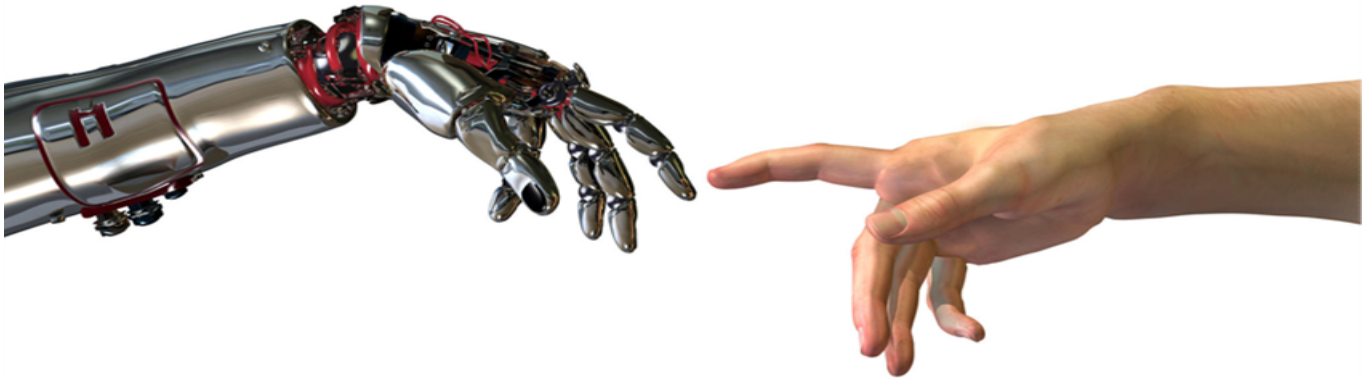
Open Questions

1) Can white lies be allowed to AI and robots?

2) Will not the increased trustworthy endanger the institution of trust in the interaction between humans and AI & Robots?

3) Is trustworthy of AI & robots a sufficient and necessary condition of trust between humans and AI & robots?

Thank you!



Nothing is stranger to man but his own image.
- Karel Čapek in *Rossum's Universal Robots* (1921)