Challenges to human dignity from developments in AI

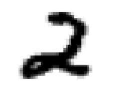
Thomas G. Dietterich
Distinguished Professor (Emeritus)
Oregon State University
Corvallis, OR USA

Outline

- ❑ What is Artificial Intelligence?
  - ❑ Near-Term Predictions for AI Technology
  - ❑ Weaknesses and Risks
- ❑ Threats to Human Dignity
  - ❑ Pervasive Surveillance
  - ❑ Injustice and Due Process
  - ❑ Simulated Empathy
  - ❑ Agency and Moral Responsibility

What is Artificial Intelligence?

- A collection of methods for creating "smart software"
  - Machine Learning
    - Give computer "training examples"
      - input → desired output



→ "2"

  - Optimization
    - Give computer an "objective function"
      - Find a delivery schedule that is as short as possible
  - Search
    - Define a "search space" and a "goal"
      - Find a sequence of moves to win in Chess

Examples of Smart Software

- Information Retrieval (web search)
- Speech Recognition (transcription)
- Language Translation
- Logistics Optimization (supply chains)
- Route Finding (driving directions)
- Drug Interaction Detection
- Drug Molecule Optimization
- Face Recognition (law enforcement, photo sorting)
- Advertisement Selection

Emerging Applications

- Smart Infrastructure (Cities; Power Grid)
- Revolution in Medical Image Analysis
- Self-Driving Cars
- (Semi-) Autonomous Weapons Systems
- Robot Care Givers
  - companions, assistants, therapists
  - sex robots?
- Human Augmentation
  - sight, hearing, smell
  - memory augmentation

Weaknesses of AI Technology

- ❑ Requires Training Data, Careful Programming, or High-Fidelity Simulation
- ❑ Expensive to Create and Maintain
- ❑ May be Difficult or Impossible to Understand
- ❑ Contains Errors and Vulnerabilities (like all software)
- ❑ Narrow and Lacks Understanding of Context

Threats to Human Dignity (1): Pervasive Surveillance

- UK and China have massive video surveillance
  - AI technology can identify and track individuals based on face and gait
  - "Activity Recognition": Jaywalking, illegal parking, littering, etc.
- Harms
  - Intrusion: Relaxation, Intimacy, Free Association
  - Errors
  - Suspicion based on correlations
  - Discrimination based on appearance, race, clothing, etc.

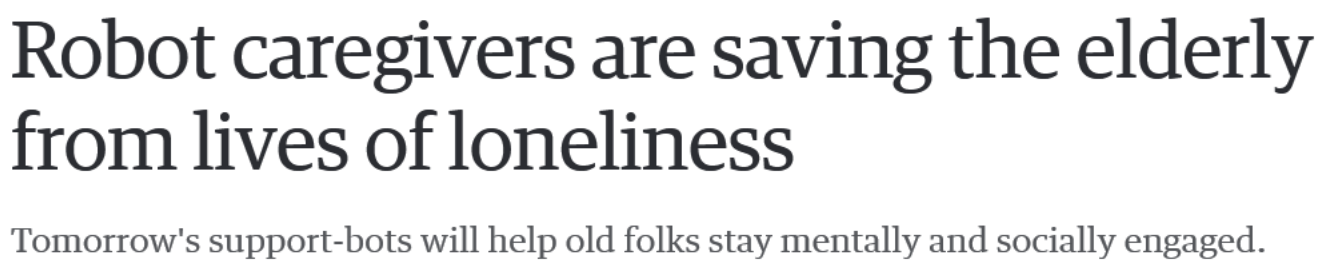
Face Recognition False Alarms

- South Wales Police
  - Average over 15 deployments: 91% (234 true alarms; 2,451 false alarms)
- Regulations
  - Requires human checking before questioning a person
  - Third party information sources must be validated prior to use
  - GDPR gives access to all data

Threats to Human Dignity (2): Injustice and Due Process

- Stop and Frisk
  - Based on face recognition, surveillance video
- US "No Fly List"
  - Criteria for inclusion are secret
  - Process for appeal is murky and slow
- China Social Credit System
  - Multiple pilot programs
  - Criteria for inclusion are published
    - Crimes
    - Failure to pay debts
    - Association with people who have low scores
  - Appeals process unclear
- Can AI-committed errors lead to false inclusion?
  - Face recognition
  - Mis-identification in financial and legal records

Threats to Human Dignity (3): Simulated Empathy



Robot caregivers are saving the elderly from lives of loneliness

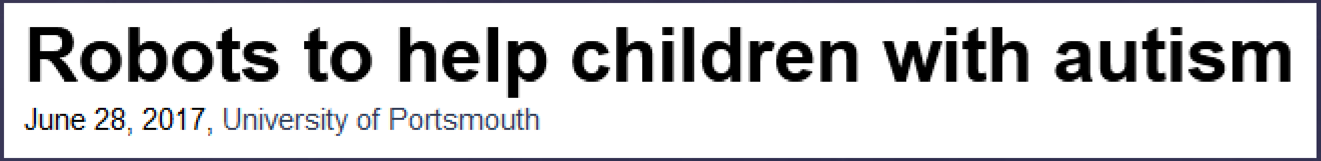
Tomorrow's support-bots will help old folks stay mentally and socially engaged.

engadget

Andrew Tarantola, @terrortola
08.29.17 in Medicine

http://www.parorobots.com/

Autism Therapy



Robots to help children with autism

June 28, 2017, University of Portsmouth

PHYS.ORG

- ❑ Children respond better to tele-operated robots than to adults
- ❑ Companies seek to automate therapy robots

Threats to Human Dignity (3): Simulated Empathy

- ❑ Computers/Robots cannot have human subjective experience
  - ❑ emotions, sensations, pain, fear
- ❑ Their understanding of human experience will always be external/behavioral
  - ❑ theories to explain human behavior
- ❑ Robot Empathy is Deception

Threats to Human Dignity (4): Moral Agency

- Under the "compatibilist" account, human decision making is deterministic and yet we hold humans morally responsible for their actions
  - provided those actions are chosen through deliberation over foreseeable consequences
- AI decision making is similar
  - AI agents evaluate the foreseeable consequences of alternative actions to choose the best action
  - AI systems created via "reinforcement learning" learn from reward and punishment
- Must we treat AI systems as morally responsible agents?

Moral Agency

- David Vladeck: Treat self-driving cars as legal persons that must carry liability insurance
  - Someone harmed by a self-driving car can sue the car and receive compensation without needing to determine which humans are responsible (operator, owner, manufacturer, software engineer, management, etc.)
  - Is this a step toward treating self-driving cars as moral agents?
  - Or is it merely an accounting trick? The insurance company decides who pays the insurance premium (operator, owner, manufacturer, etc.)

Strawsonian Approach?

- Strawson (1962) places moral responsibility in the context of social interaction
- View AI systems as incapable of genuine personal relationships and therefore not full moral agents
- Responsibility belongs to the humans who created and deployed the AI systems

Trace Responsibility Back to Humans

- Humans…
  - Formulate the AI decision making problem
  - Specify the Objective Function (the "values") of the agent
  - Collect and label the training data
  - Test and certify the safety and reliability of the AI system
  - Deploy, sell, purchase, and operate the AI system

“Machines can do many things, but they cannot create meaning. … Machines cannot tell us what we value, what choices we should make. The world we are creating is one that will have intelligent machines in it, but it is not for them. It is a world for us.”

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War.* 2018

Summary

- AI = Smart software systems
  - Existing AI systems serve as tools for human decision making
  - Future systems are likely to be more autonomous (cars, weapons systems)
- Risks to Human Dignity
  - AI-enabled attacks on freedom and human rights
    - Surveillance, Justice, and Due Process
  - Drawing a clear line between people and AI systems
    - Simulated Empathy
    - Moral Agency