# Building Ethical AI Solutions
# *Is it as hard as it seems?*

Prof Michael Rovatsos

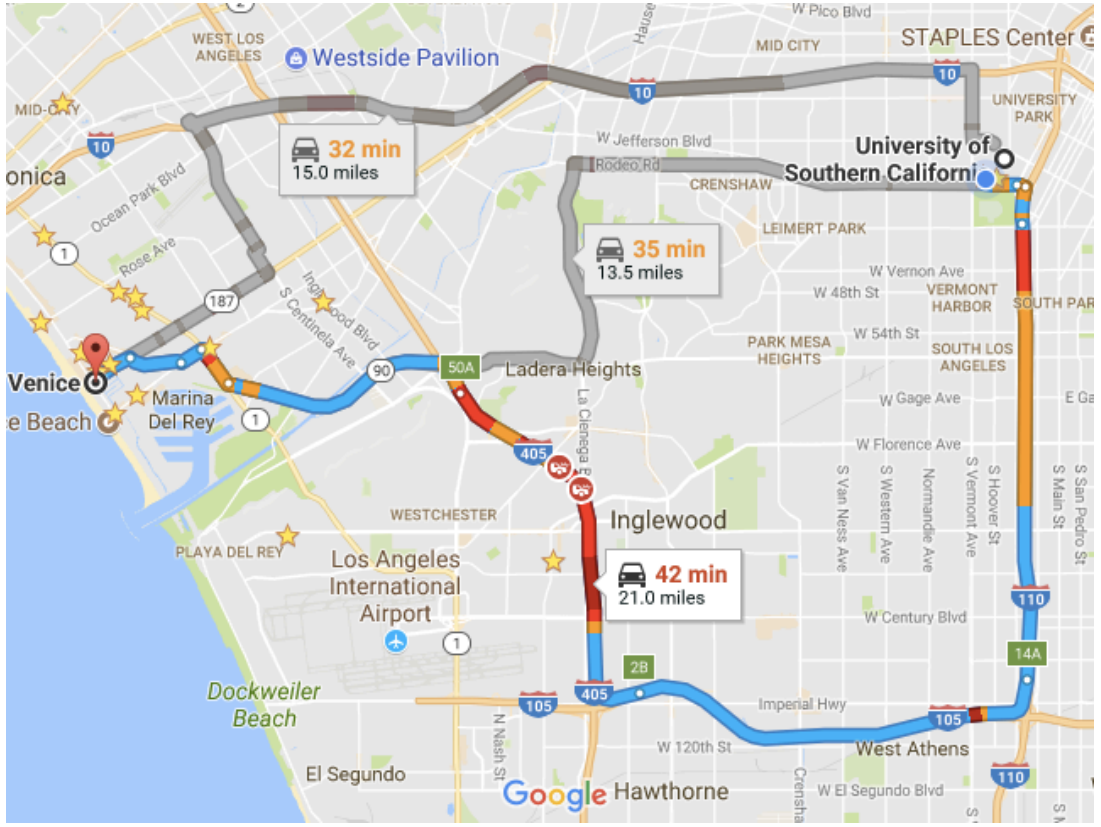
The University of Edinburgh

# Where are we headed?

- Many are concerned about the risks of AI – rightfully so
- Loss of public trust may mean we lose the potential benefits of AI
- Public debate focuses on blaming AI, rather than those responsible

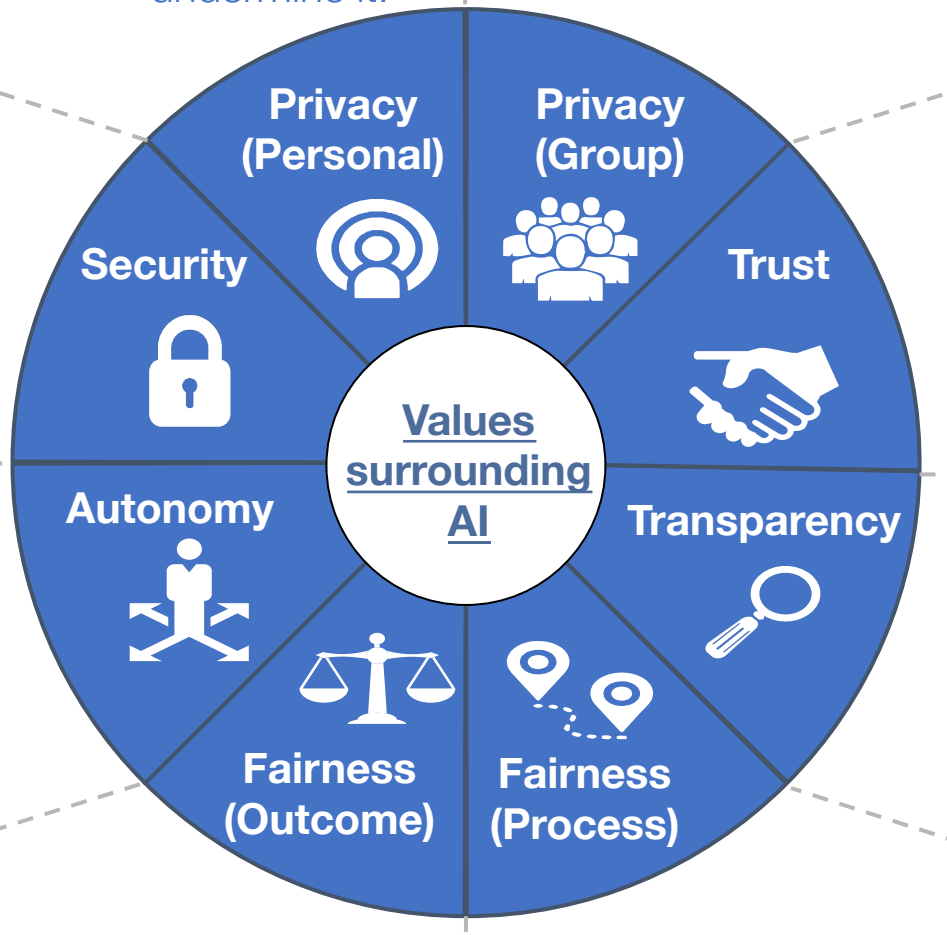
# Example: Traffic Management

- Based on mobile phone GPS tracking and incident reports
- Can reduce congestion, travel times, pollution, increase safety
- AI-based planning and route recommendation



# Ethical design questions

- Which objective function are we trying to optimise for?
- Should all users be given same weight in calculating overall welfare?
- Is the platform allowed to propose individually sub-optimal routes?
- Is it acceptable if recommendations for identical situations vary?
- Should platform take geo-advertising revenue into account?
- What parameters of the algorithm can users configure?
- Can we give paying customers preferential treatment?
- Should we “punish” deviant behavior for future routing requests?




Personal privacy is regarded as a basic human right. It is a major theme in data ethics as data driven algorithms can undermine it.

Algorithms make predictions about group level behavior: individuals are understood based on group rather than individual actions.

Actions causing avoidable (direct or indirect) threats to personal security are considered unethical.

Malicious use of data can breach security and break trust.

Mutual trust between data subjects and holders is valuable.

However, trust is *functionally* rather than *intrinsically* valuable. It is valued for its positive effects.

The ability of individual to exercise reasonable control over their actions and image.

Data and inferences drawn from data can pose threats to this self-determination.

Algorithms produce output from input. Transparency is the extent to which this process is evident or available to an individual.

It is debated whether transparency is intrinsically valuable or not.

Algorithmic outputs necessarily shape and interact with the real world. Interaction with broader systemic biases can define an algorithms ethics.

Method must respect the other intrinsic values regardless of outcome. An ethical outcome could be reached by highly unethical means.

# AI Safety Engineering

- Regulation and policy will not solve the problem of managing AI risk
- Need for solid safety engineering, and embedding it in AI tech culture
- Huge challenge for design, validation and testing, impact assessment

# The future is bright

- AI can help address many of the great challenges of our times
- We have just reached the point where it is becoming a reality
- Expecting we can have technological progress without risk is an illusion
- The best insurance policy is education!