Main Thesis / Working Hypothesis

An ethical machine can be constructed by training it towards producing reasons for its decisions that are ethically-acceptable to a human coach who undertakes the supervision of the machine.

Modern Decision-Making Systems

Okay for Ethical Decision-Making?

No! Ethical decision-making systems should be:

(P1) Judged on reasoning behind behavior.
(P2) Trained in real-life, “outside the lab”.
(P3) Supervised from counter-arguments.
(P4) Forthcoming that ethics is “subjective”.

Achieved through Machine Coaching paradigm.

Example Use of Machine Coaching

(P1) Justification of Action

- A human driver who runs over a pedestrian:
  - unintentionally because absent-minded;
  - intentionally to avoid hitting a group of children;
  - with malintent to scare pedestrian for littering;
  - with planned goal to kill pedestrian as payback.
- ML ill-fit to provide justifications to end-users.
- **Machine Coaching: reasoning over behavior.**
  - In principle, need not even make a decision.
  - Evaluation w.r.t. context, and cognitively-light.
(P2) Continuous Feedback

- Life-long learning a central tenet in our fast-changing world. Learning increasingly seen as:
  - not something that precedes other obligations;
  - not done in specialized safe places (“schools”);
  - happening in situ, at place of its eventual use.
- ML must cope with context shifts (e.g., in law).
- **Machine Coaching:** train “outside the lab”.
  - Unavailable offline gold standard for reasoning.
  - Ramifications of decisions before fully-trained.

(P3) Learning vs Coaching

- Negative reward to young child after mischief:
  - Mischiefs are wrong! OR Do not get caught!
  - even: “Don’t let me see you do that again!”. Important to explain why mischief was wrong.
- Lack of control in ML w.r.t. learned reasons.
- **Machine Coaching:** offer counter-arguments.
  - Program via justification vs learn from reward.
  - Goldilocks: cognitive / computational tradeoff.
  - Accompanying guarantees on learned reasons.

(P4) “Personalized” Ethics

- Can coach machines to reason in manner that is ethically-acceptable to their human coach.
- **Are such machines ethical? We argue yes...** ... in the same way and extent as the coach.
  - Approximately replicate the coach’s ethics.
  - No stance on who is ethical; could be a group.
- Is coach responsible for machine’s actions?
  - Are parents responsible for grown-up children?
  - Are identical twins responsible for each other?

Studies in Cognitive Systems

- *Machine-mediated diversity-aware human interactions*
  - Pilot study in 18 university and adult school sites, with 10k participants (+ non-EU).
  - Clear ethical guidance for the pilot activities and the technology development.

From Research to Innovation

- *Machine Ethics through Machine Coaching*
  - Funded by EU’s H2020 programme

(c) 2020 Loizos Michael