On the Responsible Design and Deployment of Autonomous Agents

Diederik M. Roijers **(Vrije Universiteit Amsterdam, the Netherlands & Vrije Universiteit Brussel, Brussels, Belgium)**

and

Zoltán István Zárdai **(JSPS fellow, Keio University, Tokyo, Japan)**

Workshop on new issues on technology and algorithmic ethics/privacy/fairness, related to Digital/AI/Machine-Learning

Overview

- **Focus**

- Real life, likely scenarios; technology that is with us – autonomous cars, decision support systems, security observation systems

- **Goals**

- Support some practical guiding principles for the safe design and deployment of Artificial Intelligence which is intended to operate autonomously in morally charged situations

- Offer some new arguments for why we should not let artificial **Autonomous Agents** (AAs) act without supervision except in **restricted domains**, when there are clear rules about expected behavior and consensus on what is acceptable

What we say 1

- Even if **general ethical views** (consequentialism, deontologism) are right, they do not always provide clear guidance on particular cases, they are **not handbooks**
- Operationalising general **values** like ‘happiness’ is difficult
- Even with an operational definition, maximising for these values in complex decision problems is likely intractable
- We can do better by adopting **particularism** about morality (there are objective truths, moral facts, but they do not have some easily codifiable pattern)

What we say 2

- This allows us to take morality seriously and be more **flexible** about programming at the same time, taking the rule systems of particular restricted operating domains as basis, e.g. traffic rules for cars; nursing job descriptions and codes of conduct
- AAs can have an edge over human agents due to their **reliability, speed, and immunity to performance-lowering factors** like fatigue or weakness of will
- So they can be deployed safely when their behavior can follow a **clear consensus** regarding what is the best thing to do in a certain role, there is adequate **supervision** of their activity, and their **behavior** can be **revised** and if needed they can be **decommissioned**

Particularism

- **Particularism** holds that the rationality of moral thinking and judgement does not depend on **general moral principles**
- "Moral principles are at best crutches that a morally sensitive person would not require, and indeed the use of such crutches might even lead us into moral error." Dancy, J. 2001/2017 'Moral Particularism' https://plato.stanford.edu/entries/moral-particularism/
- There is **moral truth** and there are **many morally relevant properties**
- Some are relevant in one **context** but not in another
- Example: that something is against the law might be a reason not do to it in one situation, while in another – say, during a protest – it can be a reason for doing it

General advantages of AAs over humans

- Most **humans** are **not fast and reliable** enough under pressure

- to **reason** well and go through any complex and well-balanced moral decision
- to **perform** any **complex actions** (e.g. emergency manoeuvres with a car)
- Hence, a large number of people will probably freeze or panic
- This will likely lead to **worse consequences** than an AA getting into a similar situation because even if the AA would not make a morally ideal selection, it would reliably make a **morally plausible** one it was programmed for
- Also, it could **execute** the behavior needed to carry out the selected action with a much **better chance of success** than a human agent. AAs can even opt for solutions that humans could not choose or carry out

Advantages of a particularist approach

- To make a **practical decision** human agents need to **recognise** the relevant **moral facts** which provide reasons, **weigh** them up against each other and against motivations and other reasons
- This often is a challenge even for competent human agents
- Within **restricted domains** there usually exists consensus on most moral issues and also good professional codes of conduct, clear expectations
- Cases that are contested are **known** and **documented** (euthanasia, consent)
- The number of potential **contexts** (and hence of potential moral reasons) is **limited** by the nature of the settings and duties of operations
- So, an AA which does not rely on general moral principles but rather on **domain specific information** has better chances to perform well
- When there are no clear rules humans can get involved

Relevance of trolley cases

- Trolley cases pose a dilemma: harm cannot be avoided, only minimised, but the choice is controversial
- Should it be the measure of whether an AA is ready for deployment whether it can solve **trolley problems** in a way similar or better than humans?
- **Unrealistic standard:** when humans have to make such decisions in real life they are most likely not be able to decide and act, and even when they do it is not clear why and how, and whether they do so coherently across cases
- **Trolley cases are hence not decisive. The competence of humans is not judged on its basis either and there is a real chance that AAs will do better**
- Using AAs is then less risky than letting humans operate in such situations if an AA can perform in predictable and consistent ways in **a restricted domain,** in a **safe way that complies with ethical norms**

Who makes the decisions

- **Responsibility** for the behavior of AAs: who does it rest with?
- AAs do not make **decisions** the way humans do
- They **select** actions. Action selection is the implicit result of how the reward function was designed
- It is the designing of this **reward function** that is the underlying decision, and as such, it is the designer who is the source of an AAs behavior
- When the reward function is derived from human **feedback** using some algorithm, the root cause of the agent's selections is more difficult to identify, both algorithm and feedback influence how the AA **behaves**
- However: both the creators of the algorithm, and the feedback providers are **human**

Foresight

- Can we use a particularist approach to cover every **possible moral scenario** an agent has to consider during its operation? No
- We may be able to cover most cases by identifying **a set of simple moral objectives**, and train the agent to make **trade-offs between these** different objectives according to r**eal-world examples** for which it is **clear** what the **morally right action** should be. This way the AA can learn the **consensus**, generalise and apply it sufficiently well to remain effective.
- In cases where there is no clear consensus, an AA should not make selections autonomously, but could still provide **decision support**
- When AAs are too uncertain what the consensus is, it needs defer to a responsible human

Unexpected cases

- There is also a more **difficult case**: the AA thinks it is sure enough to take a moral decision while in fact it **should not have made the selection**. If we accept the particularist view of morality this **cannot always be prevented**
- There is always a right choice in morally charged situations but there is no pattern or rule that could be applied in every case to **find** this
- The 3rd **condition** we propose follows from this: there is a responsibility to ensure that there is a process in place to **review** the behaviour of AAs
- There is a **responsibility** for the operators, its a risk they have to take
- Also, deployment needs careful planning and deliberation

Proposals

- We propose a set of **three conservative** and **cautious conditions** for when it is **permissible** to deploy AAs

1 First, there either needs to be **consensus** on what the right action is, **or** the AA needs to act in a morally **superior** way compared to humans

2 Second, there has to be adequate **supervision** of the moral selections agents make, which whenever possible will mean that non-trivial moral decisions are **deferred** to a human decision maker

3 There is a responsibility to ensure that there is a process to **review** the behaviour of AAs and in cases where it is found that AAs engage in morally wrong behavior they are **revised** to correct their behaviour, **or** get **decommissioned**

What follows

- Some examples

- Medical advice programmes – application yes, restrictions yes:

Advice on treatments, advice in diagnostics – yes

Advices or decision about risky or potentially life and death questions – no (contested normally)

- Autonomous cars – application yes, restrictions yes:

Their advantages most likely far outweigh their risks (risks from moral selection; sensors etc. are different issues)

- Airport surveillance system – applications yes, restrictions yes:

Advice to security officers - yes

Decisions on action - no

Thank you for your attention!

Keio University, 31st October 2018