Towards a Formal Ethics for Autonomous Cars

By Piotr Kulick, Robert Trypuz, and Michael Musielewicz

Presented by Samantha McDonald

Piotr Kulick

PhD in Computer Science, Philosophy

Robert Trypuz

PhD in Informatics and Telecommunications

Michael Musielewicz

PhD in Philosophy

01

Introduction

Why formalize these ethics?

- Autonomous cars use a variety of technologies to be controlled
- Implementations vary by manufacturer
- Autonomous cars a ***black box*** of functionality
  - Cars either work or they don’t
  - We have no idea how decisions are made

02

Social Acceptance of Driverless Cars

Issues of black-box decision making

The Dilemma

Benefits

* Up to 94% reduction in traffic accidents

* Increased social mobility

* Reduction in pollution

Disadvantages

* Current vehicles don’t avoid all crashes

* Transparency regarding values is unknown →

* The general public doesn’t trust black box car algorithms!

Regulation

- Regulatory bodies note the need for understanding vehicle ethics
  - US Federal Government's policy: "Even in instances in which no explicit ethical rule or preference is intended, the programming of an HAV [(highly automated vehicles)] may establish an implicit or inherent decision rule with significant ethical consequences"
  - Bundesministerium für Verkehr und digitale Infrastruktur (Germany's Federal Ministry for Digital and Transport): the very ascription of values to these objects, resting upon implicit ethical values, must be made clear so that all stakeholders can ensure that these "ethical judgments and decisions are made consciously and intentionally"
- No regulation has formalized of these ethical schemas

03

Foundations of Ethics in Cars

Can cars have ethics?

Definitions

- Agent = a system within and part of an environment which:
  - Initiates a transformation,
  - Produces an effect or,
  - Exerts power on it over time
- Interactivity = agent and environment act upon each other
- Autonomy = it is able to change state without direct response to interaction
- Adaptability = can change the transition rules by which it changes state[s]

Cars as Players

- Can autonomous cars be “players”?
  - Part of the duty-claim, liability-power, and disability-immunity relationships of an environment?
  - Beholden to give way to others – including pedestrians “who are crossing, or obviously waiting to cross at a pedestrian crossing?
- Interest theory
  - A player needs only to benefit or have an interest in the right to benefit as broadly conceived.
  - Interests can include (1) protection of passengers, (2) protection of vehicle, (3) protection of pedestrians, or (4) avoidance of crashes over breaking traffic rules
- Argument: Cars are players. As such, they should have logical schemas established to follow their interests as players in their environment

04

Ontology-based Ethical Reasoning

Framework for modeling car ethics

Ontology

The philosophical question of being, and how entities are grouped into similar categories based on relevance or importance

Described Ontologies

Environment

Types of roads

Types of places

Objects outside of car

Action

Driving actions –

Basic actions such as turn left/right

Social actions such as “save the driver”

Car Ontologies

Types of cars (passenger vs not)

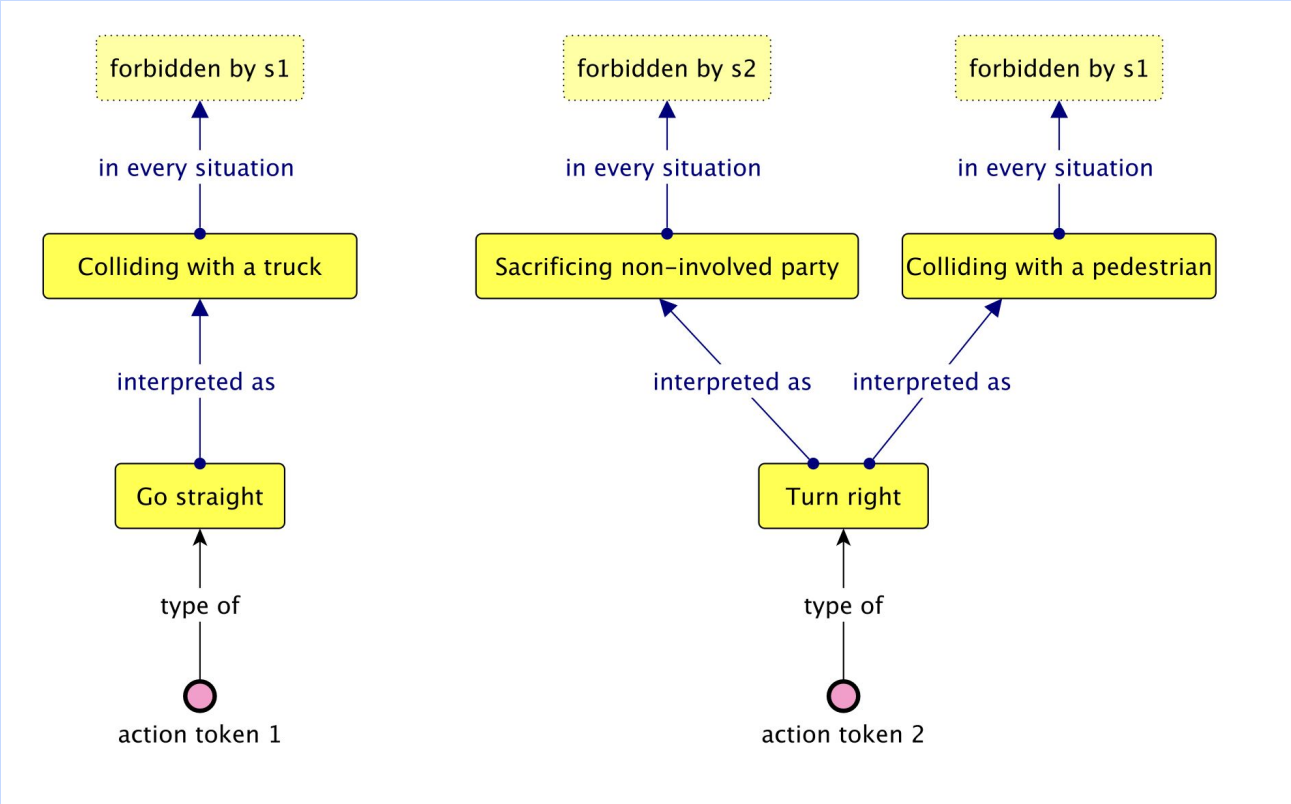
Equipment (sensors, engines, etc.)

Advance Driving System Ontology

- Self-driving car should be able to “infer driving behavior by processing knowledge”
- Semantic Web Rule Language (SWRL) rules are formulated “by means of categories taken from the ADSO Ontologies”
- Example situation: When a collision warning is activated, we have:

“Before an intersection: Give way or move forward in comply with Right-of-Way rules”, “At an intersection: Stop and give way to the other cars when upcoming collisions are detected” (see formula (1)) and “On a two-way lane: Move to the left side and give way to the other cars coming from the opposite side of the two-way lane.”

Dilemma


forbidden by s1
in every situation
Colliding with a truck
interpreted as
Go straight
type of
action token 1
forbidden by s2
in every situation
Sacrificing non-involved party
forbidden by s1
in every situation
Colliding with a pedestrian
interpreted as
interpreted as
Turn right
type of
action token 2

Addition of Ethics

- Prior examples contained some level of ethics → Forbidden actions are unethical
- Example proves there are some “impossible” decisions
- Ethical values (eg. crash into another vehicle over a pedestrian) when codified can optimize crashes when no “safe” decision is present
- Normative transparency of values allows for preference order of actions (norms) and their consequences to be clearly defined
- **No priority ranking is given

05

Conclusion

Summary of posed claims

Formalizing Ethics

- A “justification for formal ethics for autonomous cars” is presented
- Conclusion: “more powerful logical tools are needed, and we have provided a list of the basic requirements of such a logic”

- Novel ethical content was glossed over
- Vocabulary and relevant topics were well-described, but not the authors’ own ideas

06

Discussion

Further consideration points

Discussion

- As a programmer, would you feel comfortable making these ethical decisions?
- If not, who would you pass the responsibility onto? Why?

- Do you believe formalization / standardization of vehicle ethics will ever exist?
- Who will ultimately create these ethics?
  - The government?
  - A car manufacturer? A group of manufacturers?
  - Consumers / The public?


slidesgo