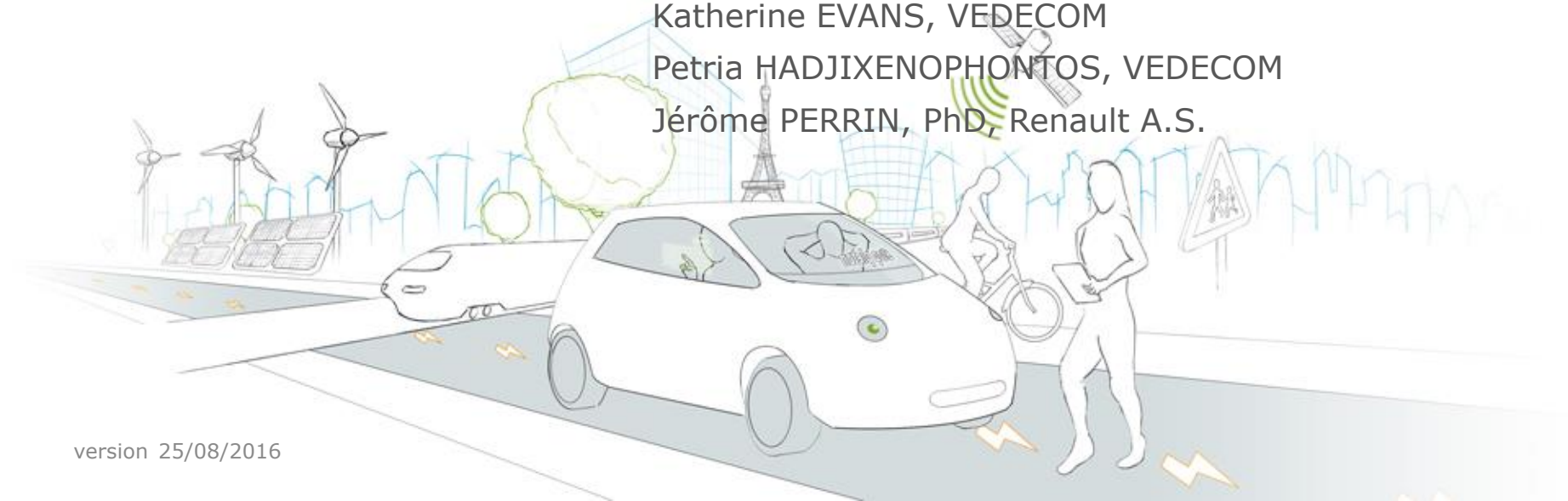
INSTITUT VEDECOM
DU VÉHICULE DÉCARBONÉ ET COMMUNICANT ET DE SA MOBILITÉ

Ethics in the design of automated vehicles: the AVEthics project

Ethics in the design of artificial agents
30 August, 2016, The Hague

Ebru DOGAN, PhD, VEDECOM
Raja CHATILA, PhD, ISIR CNRS - UPMC
Stéphane CHAUVIER, PhD, NDS – Paris Sorbonne
Katherine EVANS, VEDECOM
Petria HADJIXENOPHONTOS, VEDECOM
Jérôme PERRIN, PhD, Renault A.S.



version 25/08/2016

AVETHICS: ETHICS POLICY FOR AUTOMATED VEHICLE 2

- Funding agency: Agence National de la Recherche (ANR, France)
- Type: Young Researcher Project
- Budget: 300k
- Two PhD students

AVETHICS: ETHICS POLICY FOR AUTOMATED VEHICLE [3]

WP 1
AV ethics philosophy
VEDECOM
Iolande Vingiano
PhD student #1
UPS, SND
Stéphane Chauvier

WP 2
AV artificial ethics
VEDECOM
Mohamed Cherif-Rahal
PhD student #2
UPMC, ISIR
Raja Chatila

WP 3
AV ethics policy
VEDECOM
Ebru Burcu Doğan
PhD student #1

AVEthics

WP 4
Mangement & Communication
VEDECOM
Ebru Burcu Doğan

VEDECOM: Institute for carbon-free and communicating vehicle and its mobility

- **VEDECOM** was created as an “Institute for Energy Transition” that was established as part of the French government’s “Investment for the Future” plan. As a public-private partnership. VEDECOM is dedicated to research and training on carbon-free, sustainable individual mobility.

- The aim of VEDECOM is to facilitate and organize **pre-competitive and pre-normative research** among the main stakeholders of its three main research areas:

Vehicle electrification

Driving delegation and connectivity

Shared mobility and energy

- VEDECOM has a major role within the framework of New Industrial France project **“Automated vehicle”**.

- The institute has 10 founding members (e.g. Renault, PSA Peugeot Citroen, Valeo, IFSTTAR, UVSQ, etc.), 26 partners, and 9 associate partners.

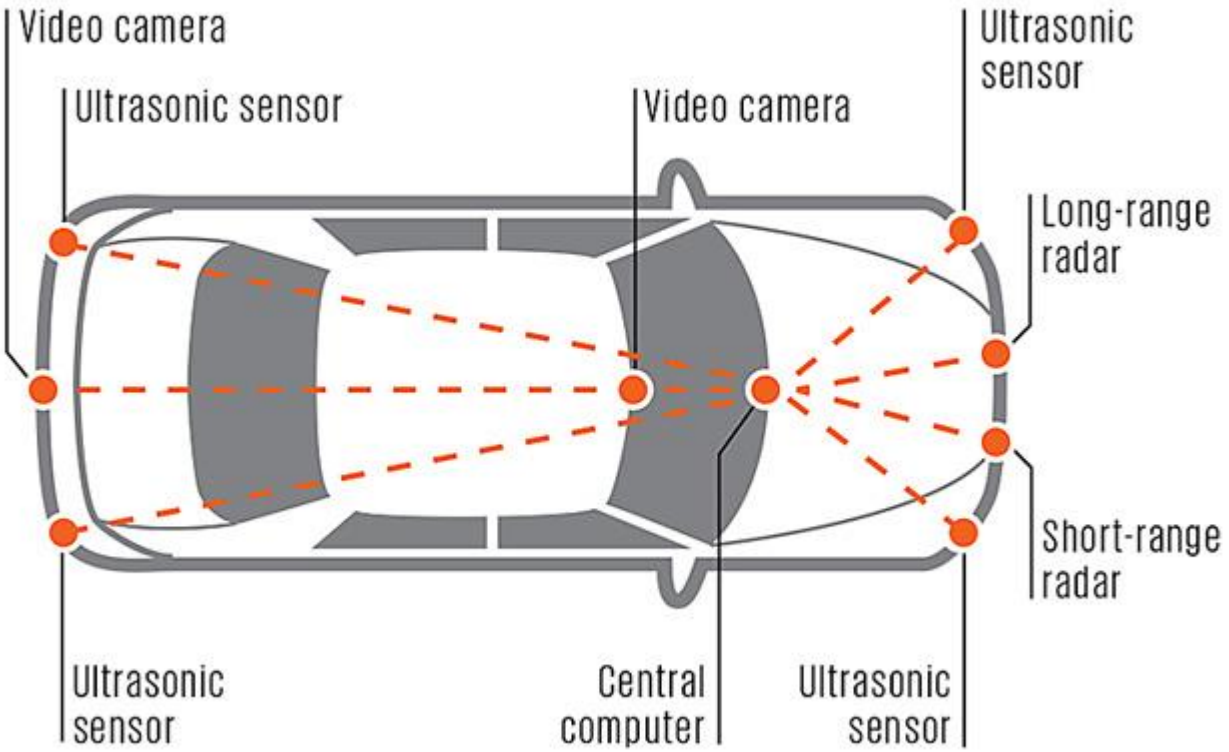
AUTOMATED VEHICLE

5

AUTOMATED VEHICLE

6

- 360 degree view enabled by sensors.
- Perception, control, and decision making algorithms



AUTOMATED VEHICLE
Why is ethical decision making relevant?

7

- AV as a social agent in a mixed traffic making decisions.
- Simple versus complex decisions
- How to handle complex decisions?
  - Human drivers have the flexibility, experience, judgmental ability to interpret the context.
  - Currently an automated vehicle makes decisions according to the Highway Code.
- Minimize risk and damage in “inevitable collision state”
- The artificial intelligence of the AV has to be able to make real time decisions of risk distribution in ethical dilemmas involving high uncertainty.
- Lin (2015) proposes that any decision involving a trade-off implies assigning a certain value to the objects, as such, ethics already becomes pertinent at the design process (i.e. control algorithms).

SPECIFICITY OF AUTOMATED VEHICLES AS ROBOTS

Contrary to personal assistant robots, or medical robots, or military robots … Automated Vehicles are characterized by three specific aspects :

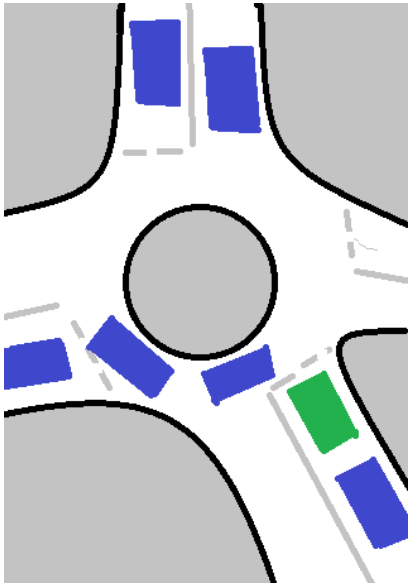
1. Used by ordinary people, not experts

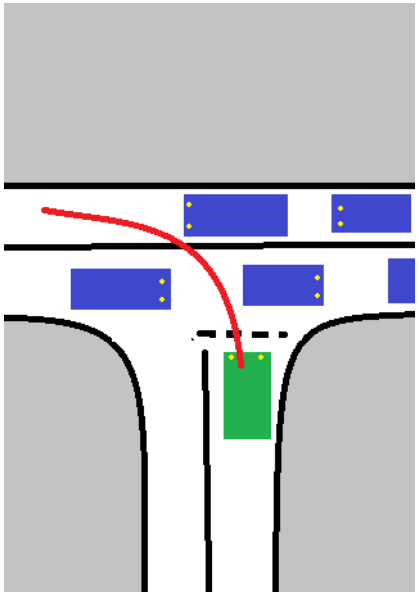
2. Co-existence of many different social agents in q complex environment (cars, pedestrians, cyclists, animals, trees, objects, infrastructures)

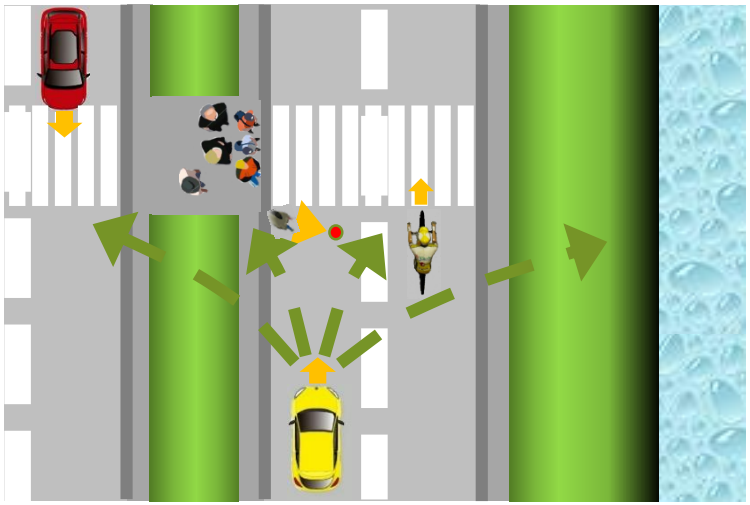
3. Implication of many stakeholders in the mobility eco-system: drivers or car occupants, other road users, car manufacturers, insurance companies, infrastructure managers, local public authorities, regulation bodies …

4. Human beings inside the robot vehicle as well as outside
   $\Rightarrow$ robot self-sacrifice is much more complex

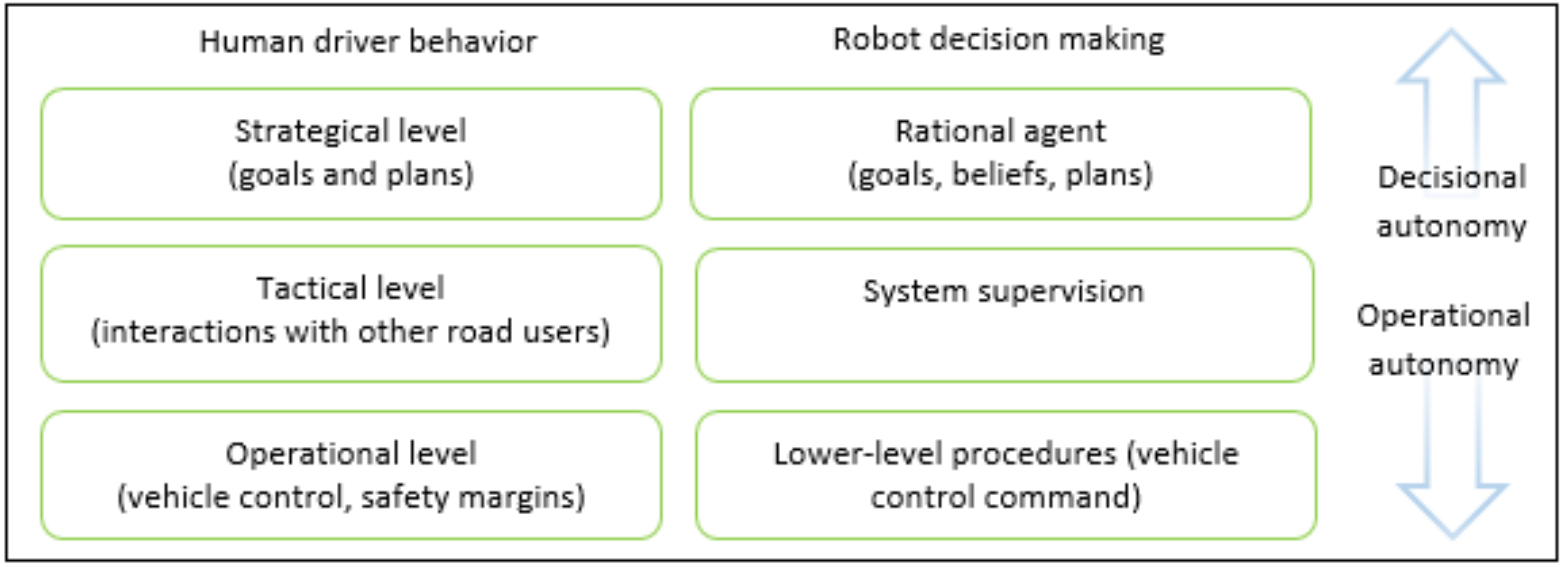
USE CASES







10



Comparison of human driver behavior and robot decision making

WHAT KIND OF ETHICS FOR AN AV?

11

- Focus on the essential needs of an artificial ethics of an AV, rather than try to implement human morality into it.

- Modular artificial moral agent (MAMA)
  - Goal oriented morality & context-specific

- In critical situations, AV ethics is a question of compromise between the **individual interests of the subject**, and those of the **public in general**

- The ethical and philosophical work behind this compromise resides in:
  - Identifying principles (values) and profiles
  - Justifying trade-offs
  - Translating formal ethics into applicable functions, hierarchies etc.
  - Categorization of entities & assigning values
  - Directing the behavior of the AV towards maximal ethical efficiency

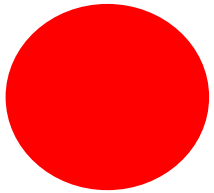
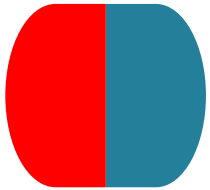
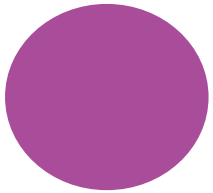
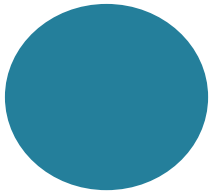
WHAT KIND OF ETHICS FOR AN AV?

12

- “Ethical valence”: to take into account the behavioral and moral basis in the *perception*
  - Numerical values assigned to (category of) every entity in the traffic environment
  - ‘Moral worth’ of the object in question, relative to the theory that is being applied.
  - Outwardly **consequentialist**: it tries *to maximize the efficiency of its action*, *through minimizing its ‘score*’ as it interacts with the traffic environment.

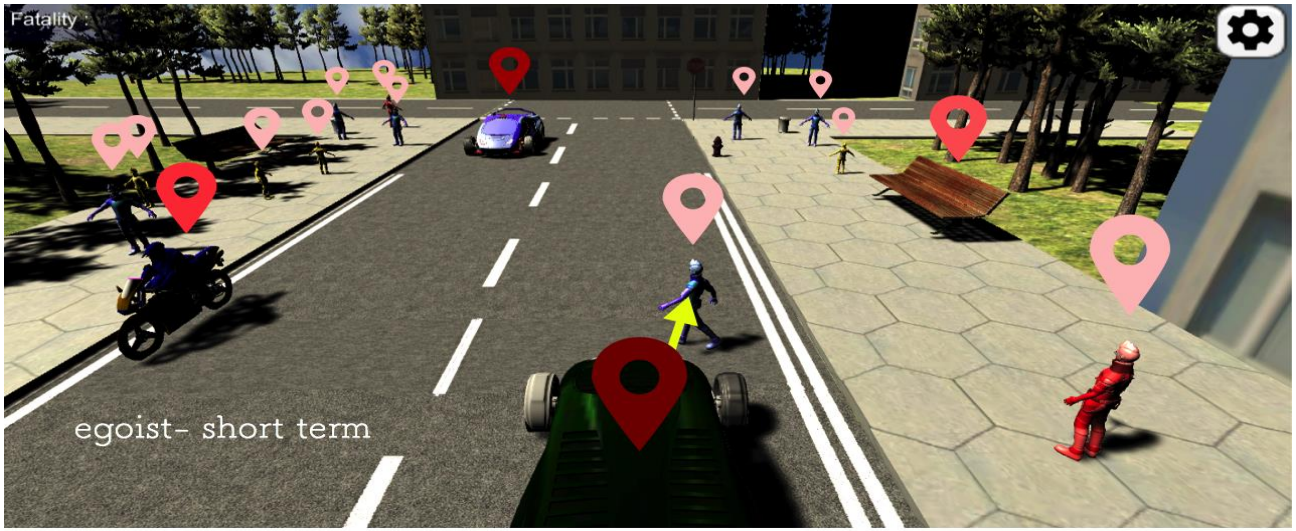
WHAT KIND OF ETHICS FOR AN AV?

13

EGOIST	RISK AVERSE ALTRUIST	SUFFICIENT ALTRUIST	PURE ALTRUIST
			
"I"	"You & I"	"We"	"You"
Welfare maximizing (risk minimizing) for AV subject, objective preferences are not included in calculation	Welfare maximizing for society, but limited by a *no-harm principle*: the AV subject can never suffer a fatal accident	Welfare maximizing for the most vulnerable entity in the environment (including the AV subject)	Welfare maximizing for society, favoring the most vulnerable entity, the AV subject is not included in calculation

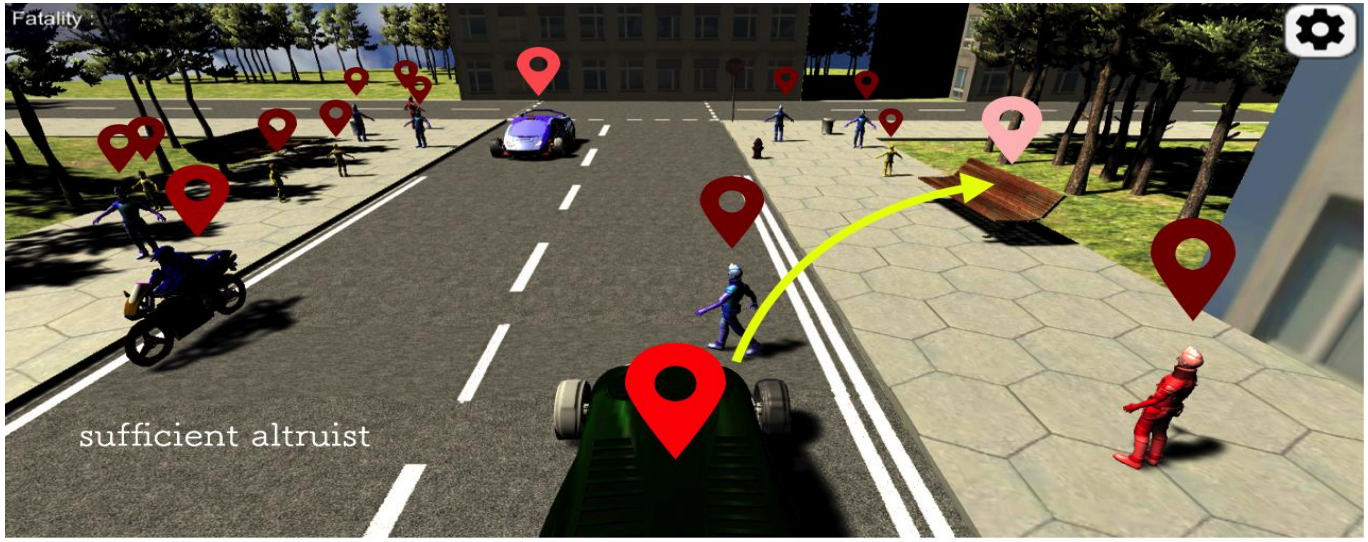
WHAT KIND OF ETHICS FOR AN AV?

14



WHAT KIND OF ETHICS FOR AN AV?

15



INFORMATICS

16

- Ethics principles translation into algorithms in a traceable way
- Quantify sensor data
- Categorization of entities
  - Uncertainty in categorization
  - Uncertainty in action implementation
- Path generation in dynamic environment

INFORMATICS

17

- Game interface
  - Utility calculations and moral choices based on the profile

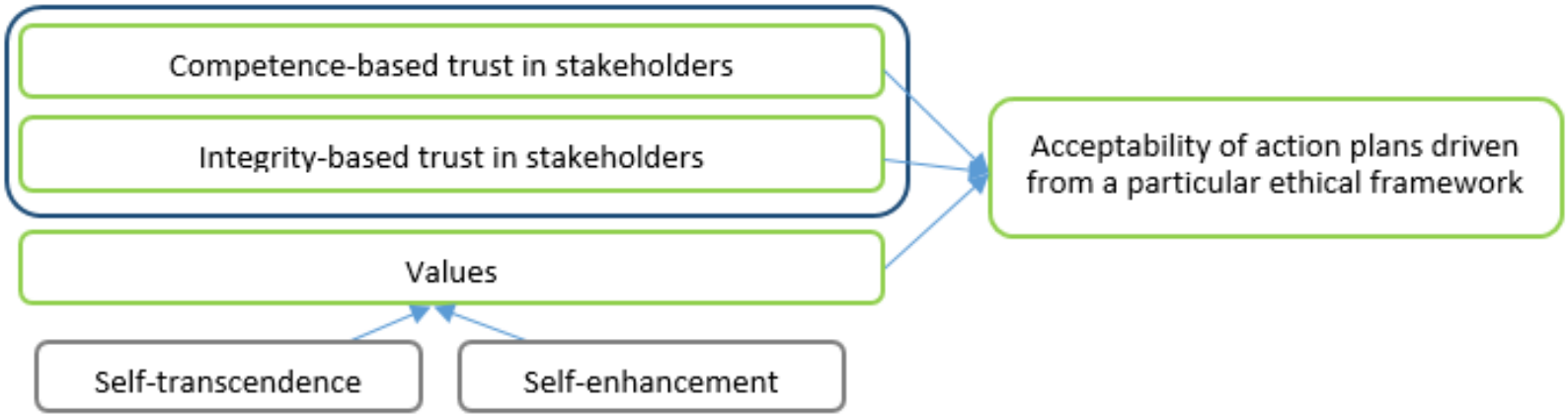
ACCEPTABILITY

18

- If AV would reduce the accidents, why do we worry about its acceptability?
    - Lay people's decision making based on qualitative notion (e.g. controllable, voluntary, familiar)
    - Under-valued risk: high probability of a small effect (e.g. traffic accident)
    - Over-valued risk: low probability of a large effect (e.g. nuclear)
- Deciding for yourself versus letting the vehicle decide for you
- Human *reaction* (immediate) versus vehicle *decision* (forethought) (Lin, 2015)
- Blame the software: People are more likely to accept a higher chance of injury by a human driver than by a software.

ACCEPTABILITY

19



ACCEPTABILITY

20

- Human ethical decision-making is a mix of emotions and reason.
  - Ignore objective information and its legitimacy

- Values: initial judgments of acceptability are based on values.
  - Collective outcomes versus personal outcomes, which might be in conflict in controversial issues.

- Trust… very strongly and positively related to acceptability.
  - Competence-based versus integrity-based
  - The trust-acceptability relationship is not straightforward for issues of high-moral importance.

ACCEPTABILITY

21

- Three steps:
  - Focus group: stakeholders' opinions about the AV ethics and use cases
  - Survey study: explication interviews based on video recording of the game interface
  - Behavioral study: simulator study integrating the game interface

CONCLUSIONS

22

- AVs will have to make ethical decisions in accident situations
  - Transgression of the Highway Code
- Consensus on ethical principles and validation cases
- How to implement ethics in an AV?
  - Machine learning versus programming
- Acceptability of the ethical principles for the end user as well as the main stakeholders
- Dynamic process: penetration rate of AV

INSTITUT VEDECOM
DU VÉHICULE DÉCARBONÉ ET COMMUNICANT ET DE SA MOBILITÉ

Thank you for your attention

ebru.dogan@vedecom.fr