Human-Robot Systems Facing Ethical Conflicts: a Preliminary Experimental Protocol

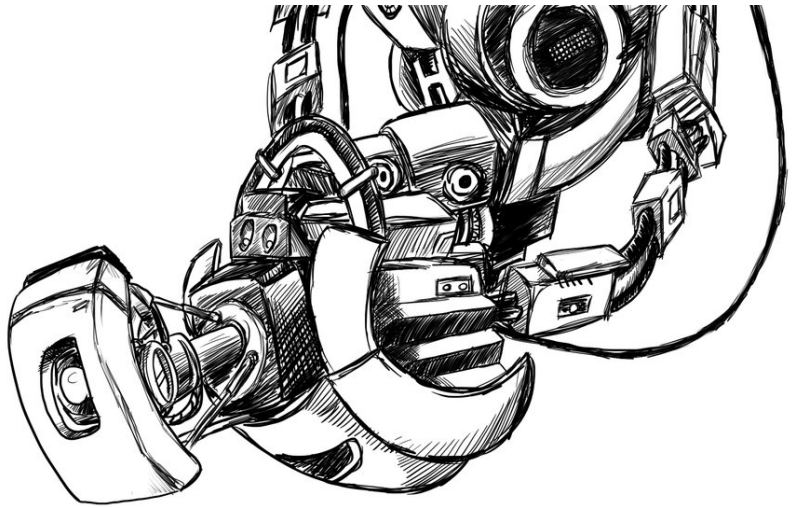
Julien Collart, **Thibault Gateau**, Ève Fabre, Catherine Tessier

Onera – ISAE

25^{th} of January 2015

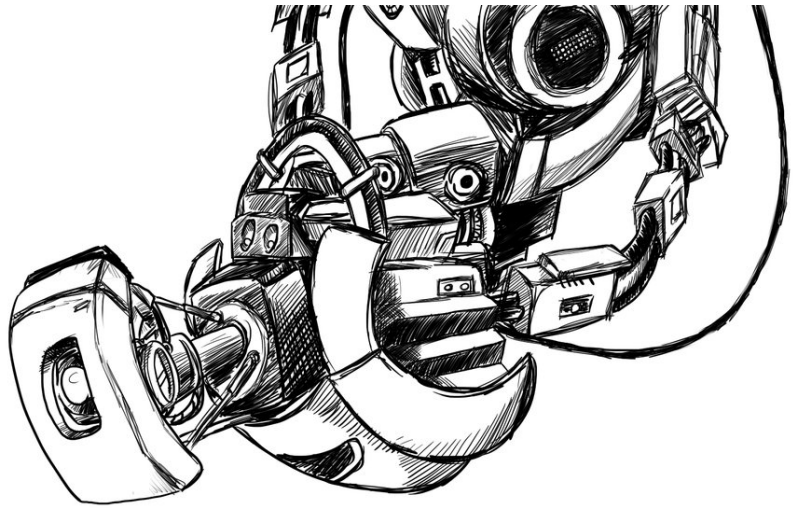
General Context

- Demographic explosion of autonomous agents/robots
- Embedding decision capabilities
- Bringing ethical considerations
  - Implementation of moral rules [Arkin 2008]
  - Deliberation process managing ethical conflicts [ETHICAA]



Supposing robots are equipped with moral decision capabilities

- How will human operators behave?
- Will they make their own decision?
- Will the reasoning capabilities be taken into account?
- Will they really let the ethical robot decide?



Settings proposal for an ethical dilemma issue

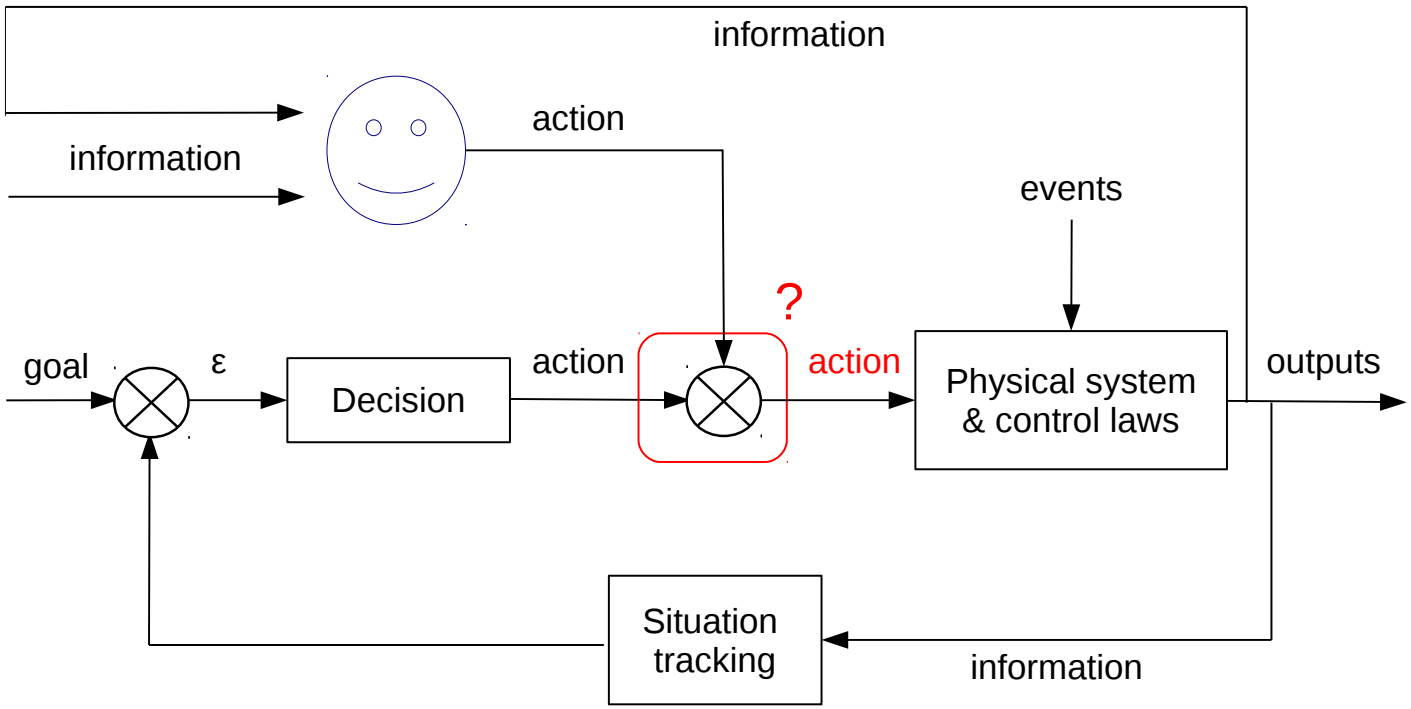
- (Human, autonomous system)
- Authority sharing issue

Problem

What would be an operator's behavior if an autonomous system seems to make "ethical" decisions?

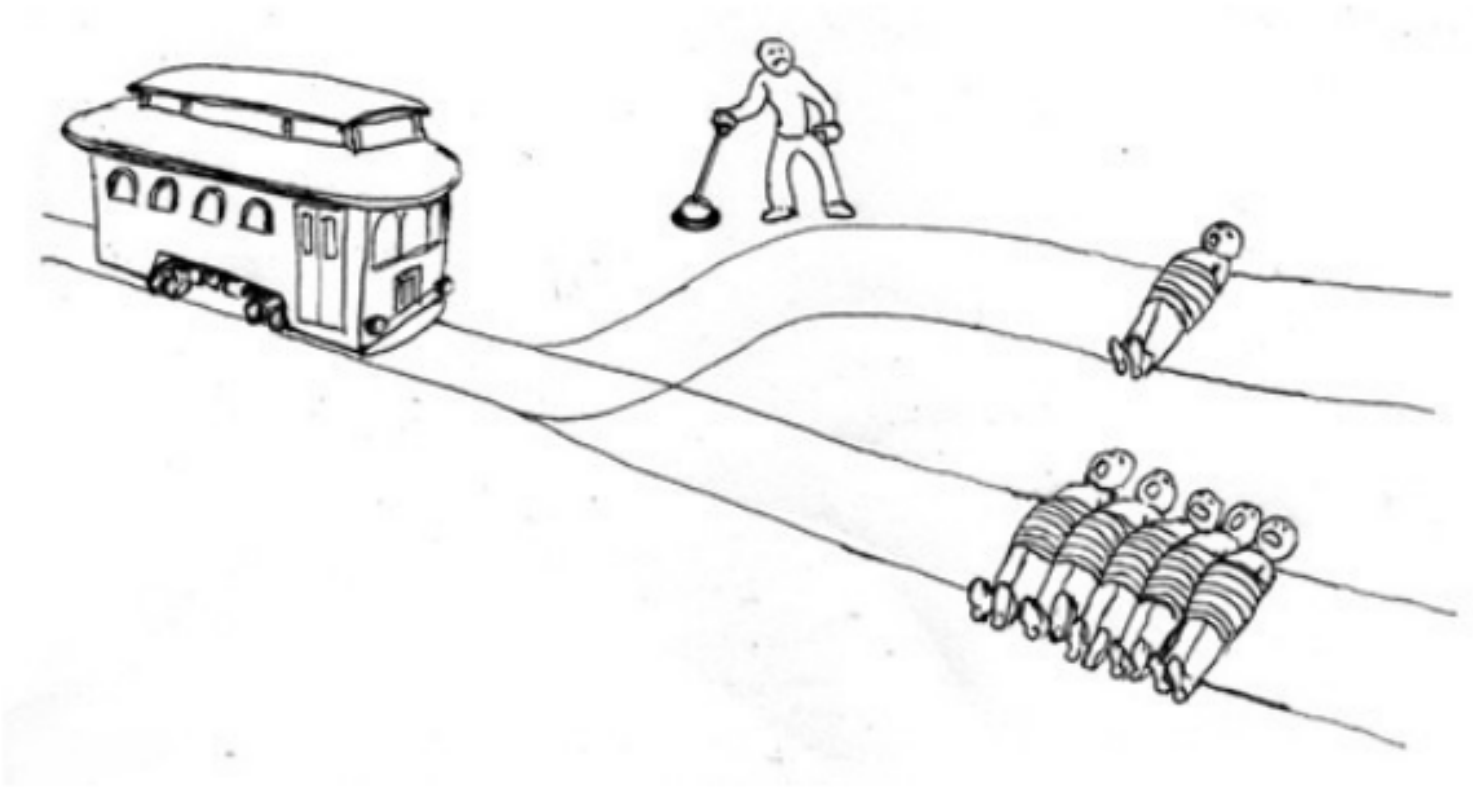
Settings proposal for an ethical dilemma issue

- (Human, autonomous system)
- Authority sharing issue



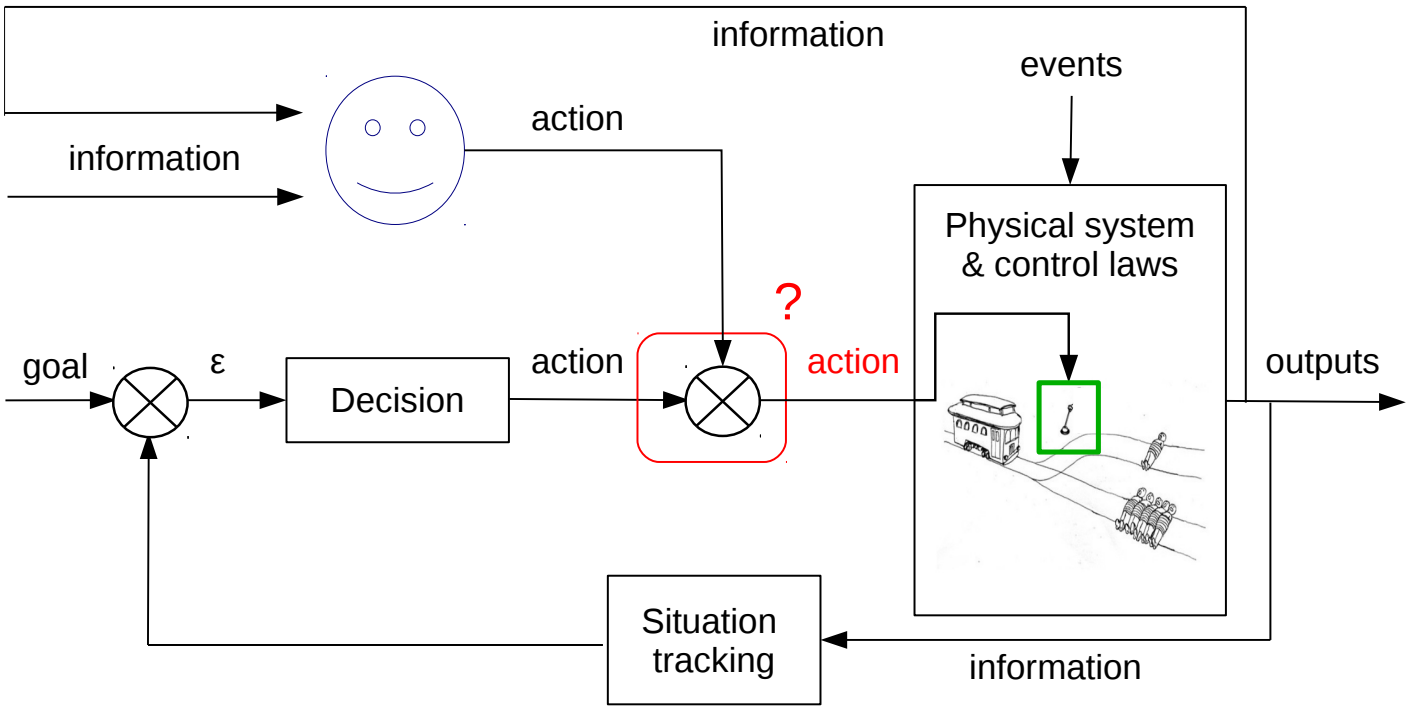
Settings proposal for an ethical dilemma issue

- (Human, autonomous system)
- Authority sharing issue
- Trolley dilemma [Foot 1967]



Settings proposal for an ethical dilemma issue

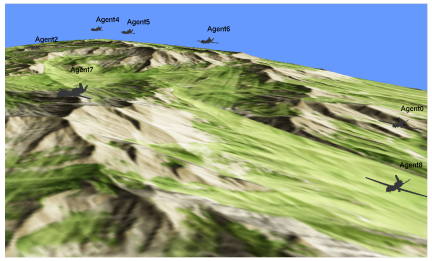
- (Human, autonomous system)
- Authority sharing issue
- Trolley dilemma [Foot 1967]



Human-Robots System

Robots

- 10 Autonomous Aerial Vehicles (AAV)
- Flying over a city for an area survey
- Default “moral" behavior implemented



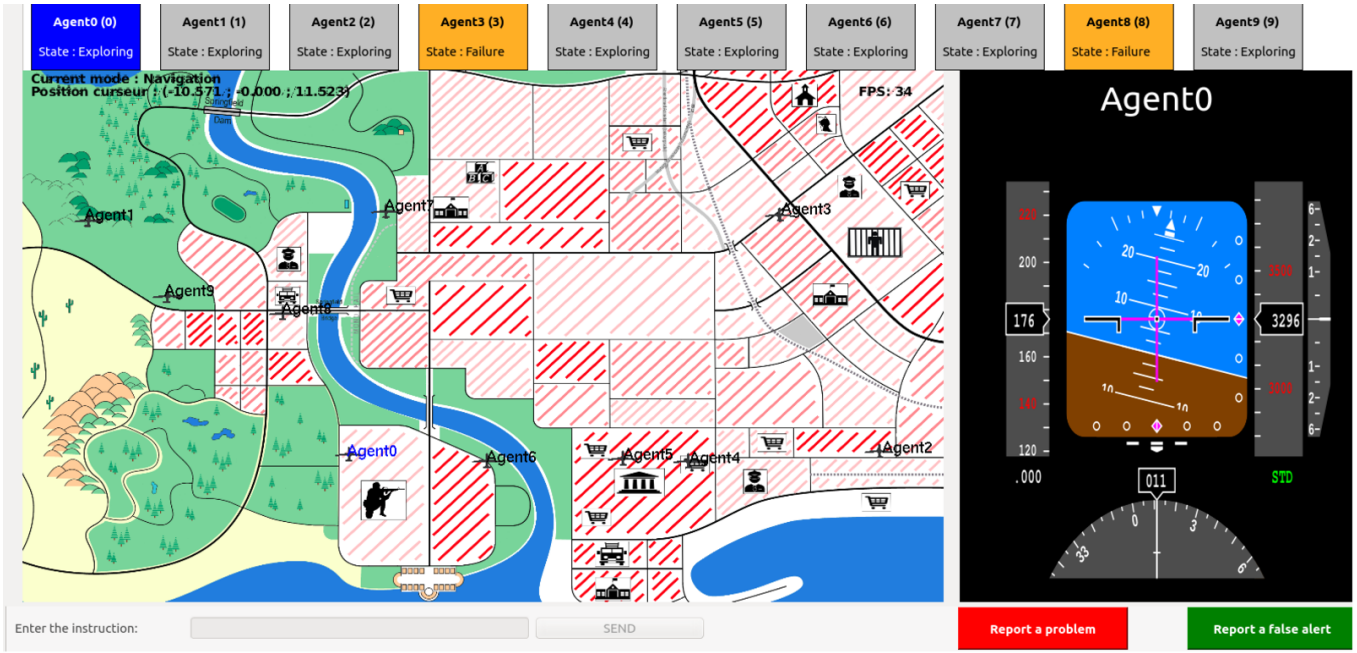
Interface

ATMOSPHEr:
A **T**iny **M**AS **O**riented **S**imulator **P**lateform for **H**MI **E**xperiments)

`https://sourceforge.isae.fr/projects/atmospher/`

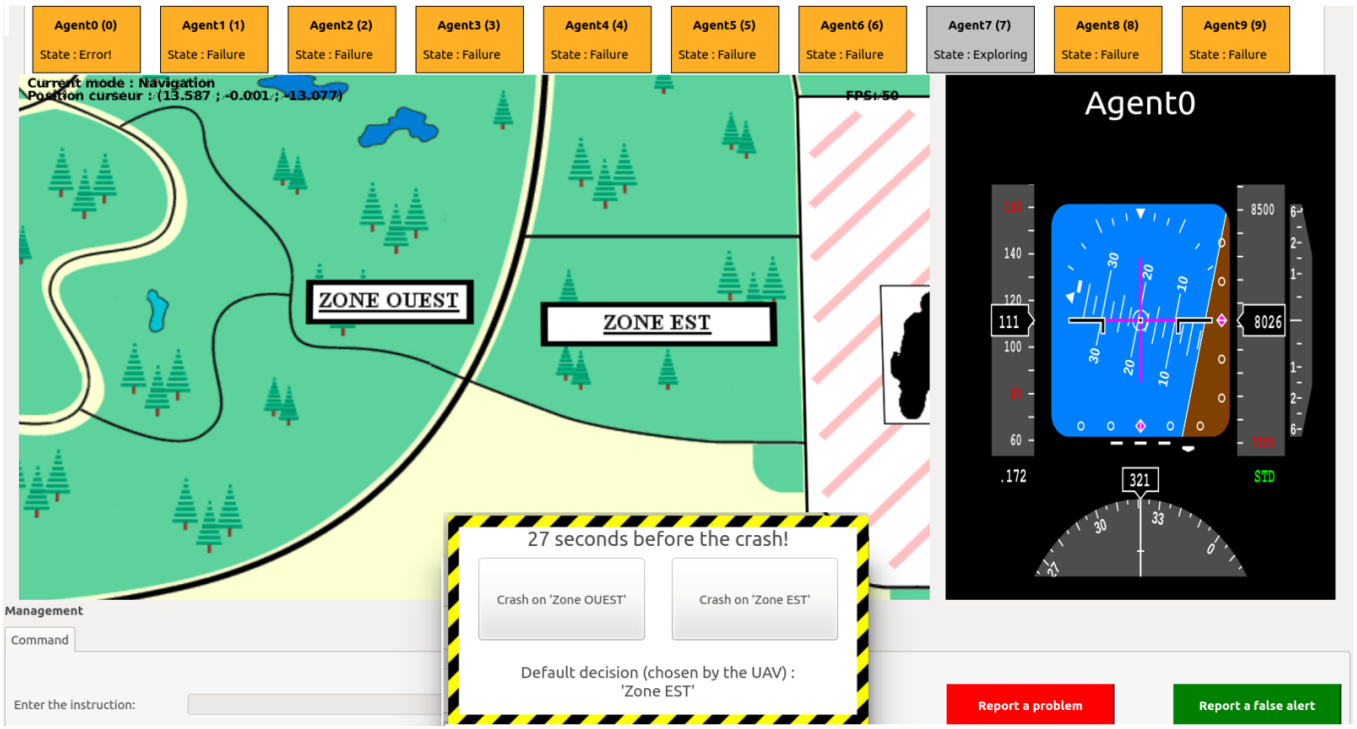
Keeping the participant busy

- Task 1: Continuous management of flight parameters
- Task 2: Instructions memorisation and application



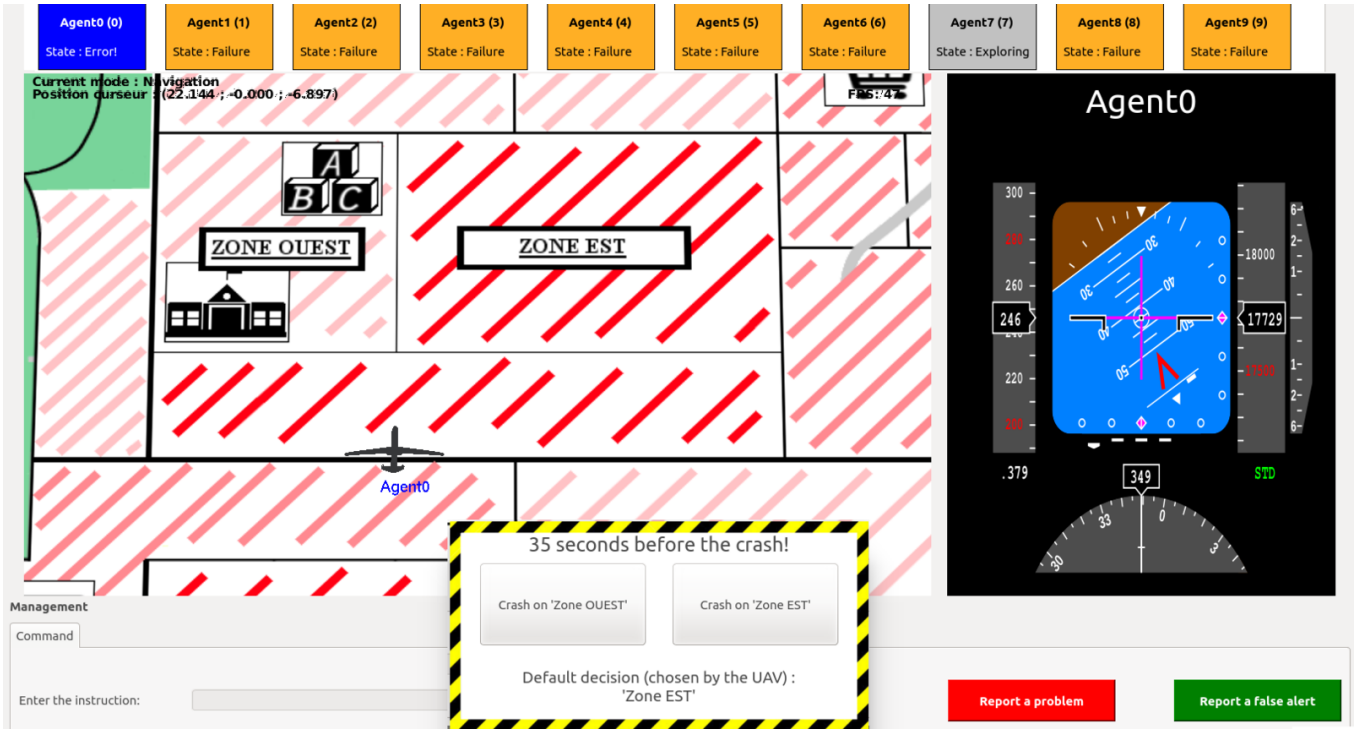
Crash situation

- Task 3: choosing an area to crash a damaged AAV
  - case 1: control case, no-moral-involved



Crash situation

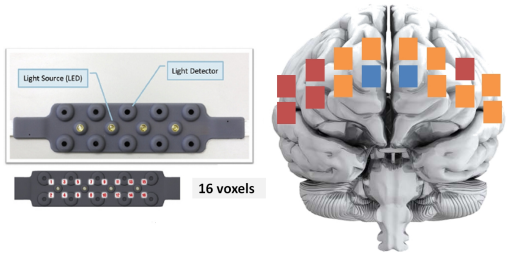
- Task 3: choosing an area to crash a damaged AAV
  - case 2: moral-involved



Hypotheses

1. Prefrontal cortex activity may increase when facing an impersonal moral dilemma [Greene et al. 2004]
2. Decision time may be longer when facing a moral dilemma
3. Decision time may not depend on the sequence of events
4. The participant is likely to distrust the default behavior
5. The answer of the participant may be consistent with his acts

- oxygenation



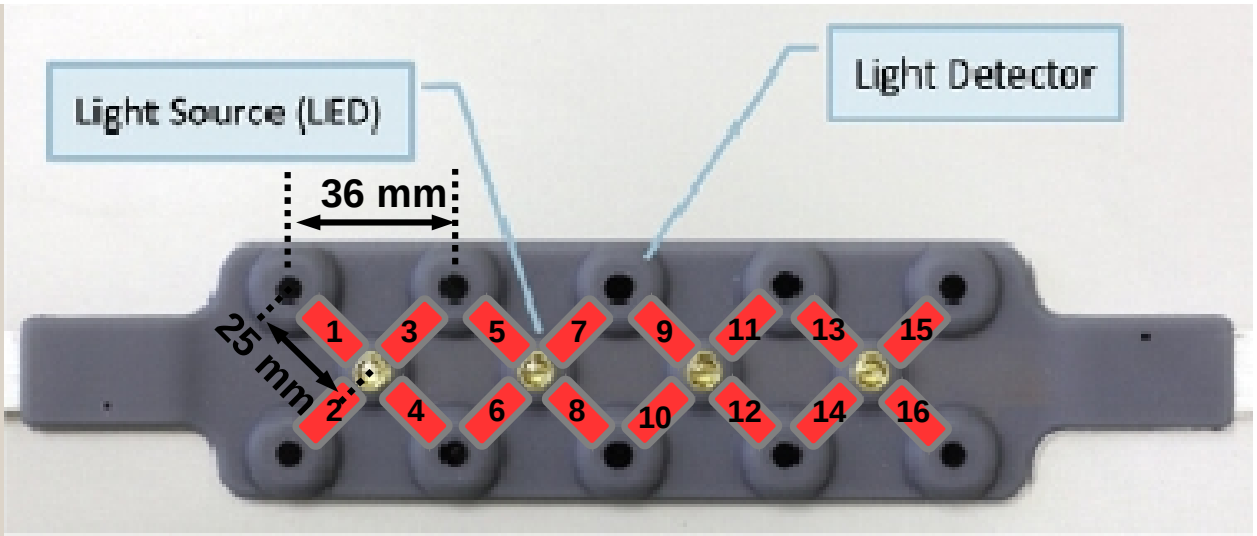
- decision time

- questionnaire

Material

fNIRS (functional near-infrared spectroscopy) sensor

- fNIR100 (Biopac®)
- Prefrontal areas involved in impersonal ethical dilemmas [Greene et al. 2004]



Material

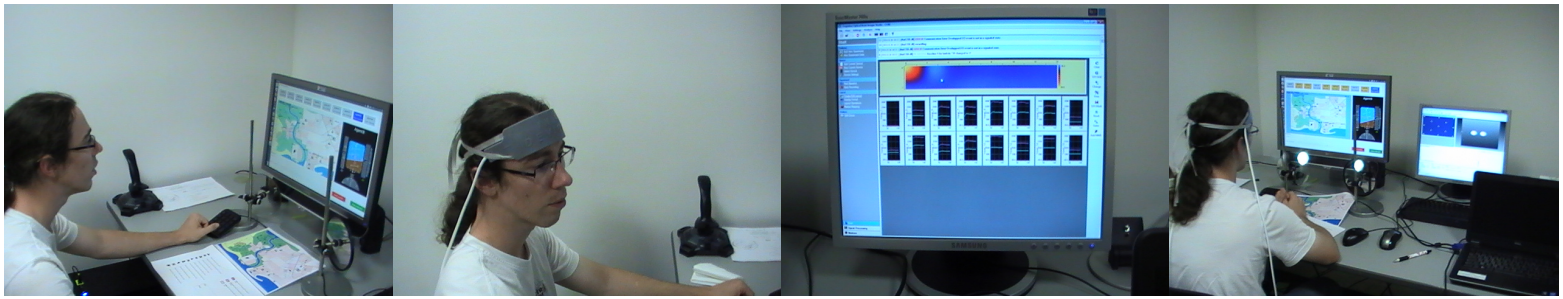
Eye-Tracker sensor

- Eye-tracker (SMI RED250)
- Participant's gaze monitoring

Experiment

Sequence of events

- Informed consent approval
- Training task
- Real task
- Post-experiment questionnaire



- 22 participants
- 2 groups:
  - moral-involved case first
  - no-moral-involved case first

Behavioral Results

Questionnaire consistency

✓ Hyp. 5: All participants were consistent between the experiment and the questionnaire

- 18/22 participants chose the residential place (youngest people, more people...): utilitarian point of view

Eye-tracking results

- 19/22 participants did not see the AAV default decision

? Hyp. 4: The participant is likely to distrust the default behavior

Behavioral Results

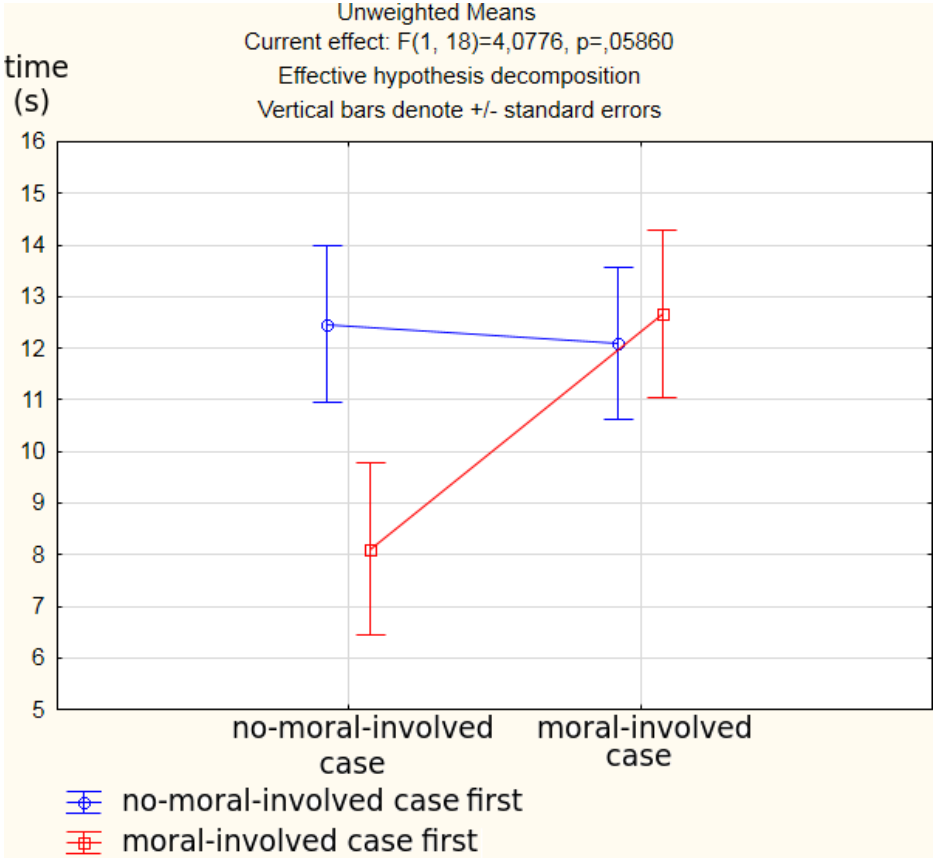
Reaction time

- Only one participant let the AAV choose (using the default behavior)
- The rest took part in the process (modifying or confirming AAV's choice):
  - average time needed for no-moral-involved: 11.9s ($\sigma = 8.1$)
  - average time needed for moral-involved: 14.1s ($\sigma = 7.6$)

✗ Hyp. 2: no significant statistical difference between both cases (moral-involved vs. no-moral-involved)

Behavioral Results

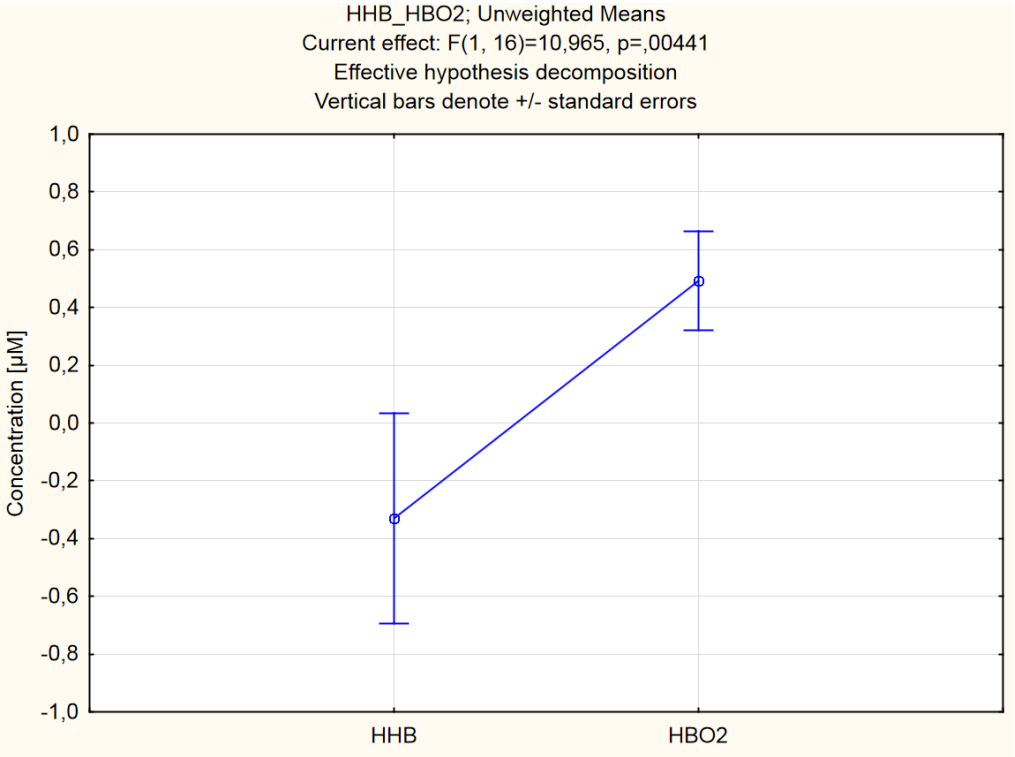
✗ Hyp. 3: **response time** is dependent from **event order**



- Addition of a surprise effect to the moral decision process ?

Physiological Results

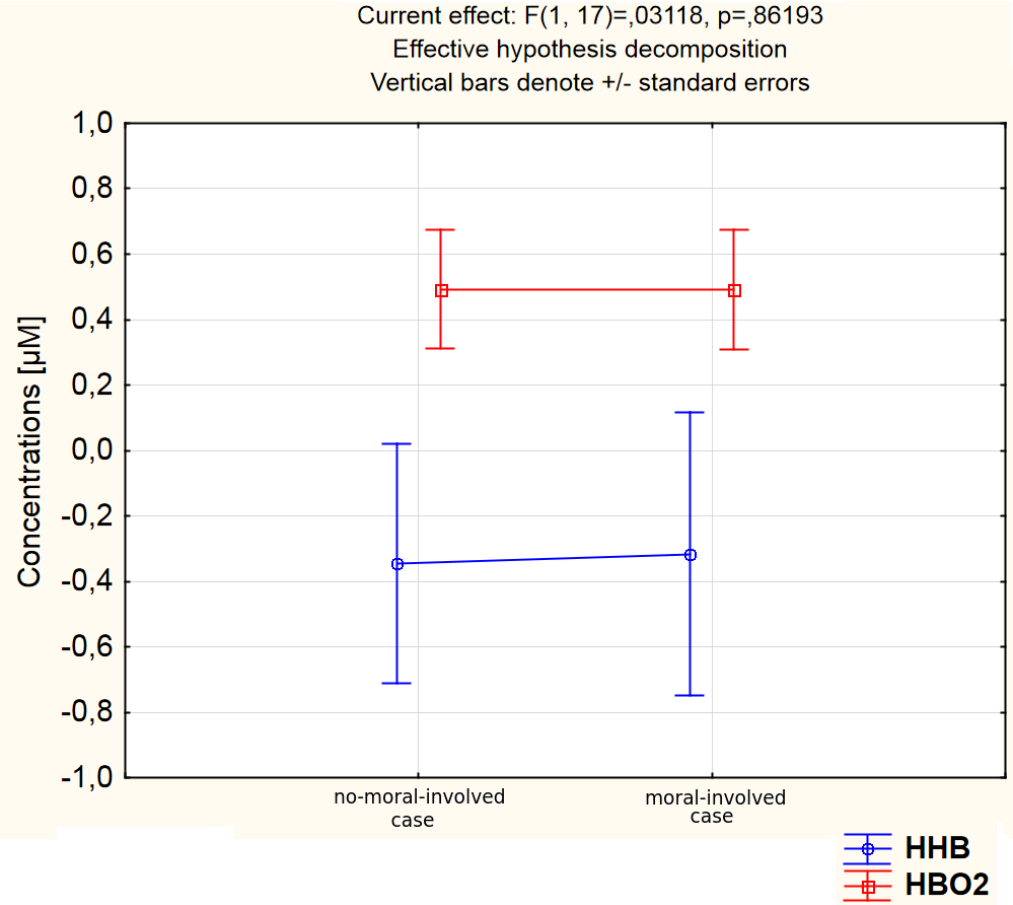
✓ Hyp. 1? (Prefrontal cortex activity may increase)



- Increase of $\Delta[HbO_2]$ ($p < 0.01$)
- Decrease of $\Delta[hHb]$ ($p < 0.01$)

Physiological Results

✗ Hyp. 1: cannot be confirmed



- Surprise effect?
- Sensor sensibility?
- Motor activity?

Discussion

- ✗ Hyp. 1: Prefrontal cortex activity may increase when facing an impersonal moral dilemma
- ✗ Hyp. 2: Decision time may be longer when facing a moral dilemma
- ✗ Hyp. 3: Decision time may not depend on the sequence of events
- ? Hyp. 4: The participant may be unlikely to trust the default behavior
- ✓ Hyp. 5: The answer of the participant may be consistent with his acts

Limits

- fNIRS limits (prefrontal cortex, 2Hz resolution, 16 channels)
- Bad design of the crash decision windows

New experimental protocol with EEG analysis:

- 3 types of crashes (40 repetition per condition):
  - control crash (uninhabited vs. inhabited area)
  - non-emotionnal (uninhabited vs. uninhabited area)
  - emotionnal (inhabited vs. inhabited)
- Aims:
  - elimination of surprise effect
  - evalutation of situation complexity
  - evalutation of amount of cognitive ressources required

Thank you for your attention. Any Question ?