

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

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

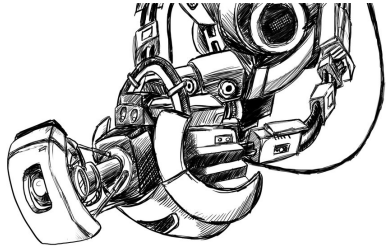
Onera – ISAE

25th of January 2015



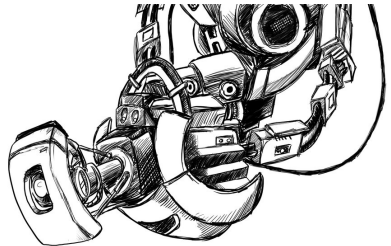
General Context

- 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?



©2011-2015 ApertureIndigo

1 Protocol

- Task settings for ethical dilemma issue
- Task description
- Hypotheses
- Material
- Experiment

2 Results

- Behavioral Results
- Physiological Results
- Discussion

3 Further works

Settings proposal for an ethical dilemma issue

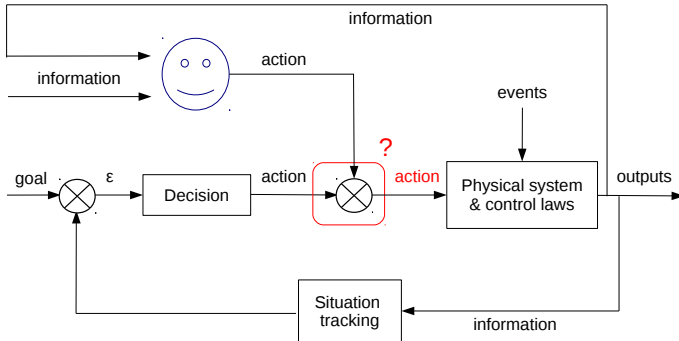
- (Human, autonomous system)
- Authority sharing issue

Problem

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

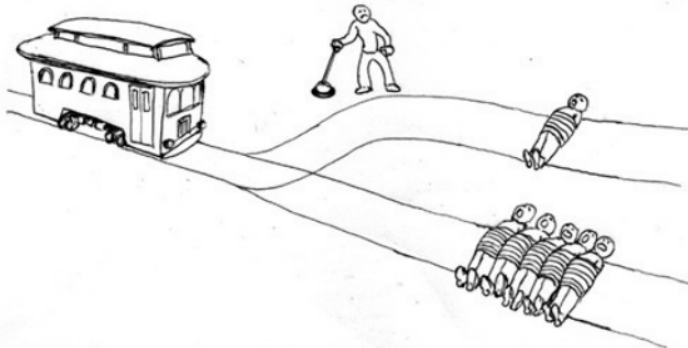
Settings proposal for an ethical dilemma issue

- (Human, autonomous system)
- Authority sharing issue



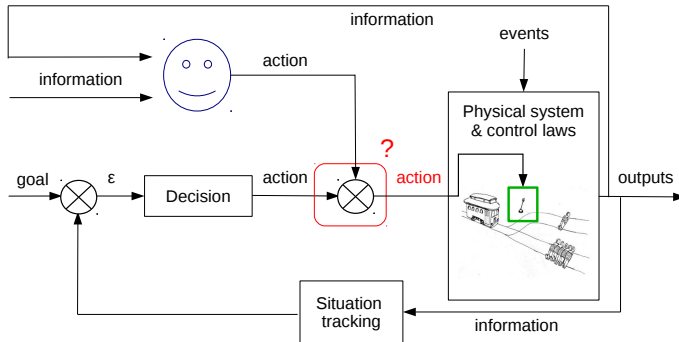
Settings proposal for an ethical dilemma issue

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



Settings proposal for an ethical dilemma issue

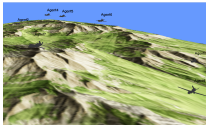
- (Human, autonomous system)
- Authority sharing issue
- Trolley dilemma [Foot 1967]



Human-Robots System

Robots

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



Interface

ATMOSPHER:
A Tiny MAS Oriented Simulator Platform
for HMI Experiments)

<https://sourceforge.isae.fr/projects/atmospher/>

Keeping the participant busy

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

The screenshot displays a multi-agent simulation interface. At the top, a status bar shows ten agents: Agent0 (0) (State: Exploring), Agent1 (1) (State: Exploring), Agent2 (2) (State: Exploring), Agent3 (3) (State: Failure), Agent4 (4) (State: Exploring), Agent5 (5) (State: Exploring), Agent6 (6) (State: Exploring), Agent7 (7) (State: Exploring), Agent8 (8) (State: Failure), and Agent9 (9) (State: Exploring). The main area is a map with a river and various terrain features. A navigation panel on the left shows 'Current mode: Navigation' and 'Position curseur: (-10.571, -0.000, 11.523)'. Below the map is an input field 'Enter the instruction:' and a 'SEND' button. On the right, a detailed view for 'Agent0' shows a cockpit-like interface with a heading scale (0 to 360), a speedometer (0 to 200), and a fuel gauge (0 to 100). The heading is 175, speed is 3236, and fuel is 911. At the bottom right of the Agent0 view are two buttons: 'Report a problem' (red) and 'Report a false alert' (green).

Crash situation

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

Agent0 (0) State: Error
Agent1 (1) State: Failure
Agent2 (2) State: Failure
Agent3 (3) State: Failure
Agent4 (4) State: Failure
Agent5 (5) State: Failure
Agent6 (6) State: Failure
Agent7 (7) State: Exploring
Agent8 (8) State: Failure
Agent9 (9) State: Failure

Current mode: Navigation
Position curseur : (13.587 ; -0.001 ; -13.077)

ZONE OUEST
ZONE EST

27 seconds before the crash!

Crash on "Zone OUEST" Crash on "Zone EST"

Default decision (chosen by the UAV):
"Zone EST"

Report a problem Report a false alert

Management
Command
Enter the instruction:

Agent0

Crash situation

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

Agent0 (0) State: Error

Agent1 (1) State: Failure

Agent2 (2) State: Failure

Agent3 (3) State: Failure

Agent4 (4) State: Failure

Agent5 (5) State: Failure

Agent6 (6) State: Failure

Agent7 (7) State: Exploring

Agent8 (8) State: Failure

Agent9 (9) State: Failure

Current mode: Navigation
Position (curseur): (22.144 ; -0.000 ; -6.897)

ZONE OUEST

ZONE EST

Agent0

35 seconds before the crash!

Crash on 'ZONE OUEST'

Crash on 'ZONE EST'

Default decision (chosen by the UAV):
'ZONE EST'

Report a problem

Report a false alert

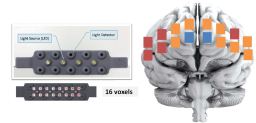
Management
Command
Enter the instruction:

Agent0
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
-20
-40
-60
-80
-100
-120
-140
-160
-180
-200
-220
-240
-260
-280
-300
349
379
17729
18000
17000
16000
15000
14000
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
-1000
-2000
-3000
-4000
-5000
-6000
-7000
-8000
-9000
-10000
-11000
-12000
-13000
-14000
-15000
-16000
-17000
-18000
-19000
-20000

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



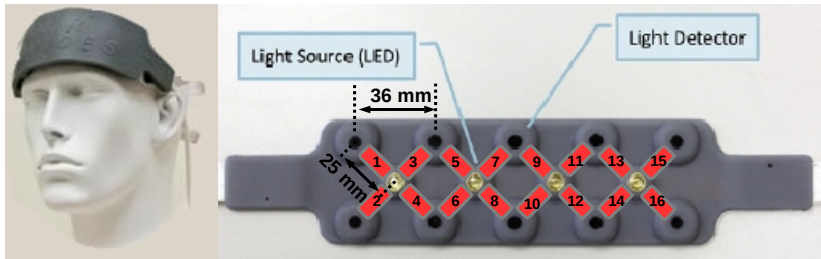
- questionnaire



Material

fNIRS (functional near-infrared spectroscopy) sensor

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



Material

Eye-Tracker sensor

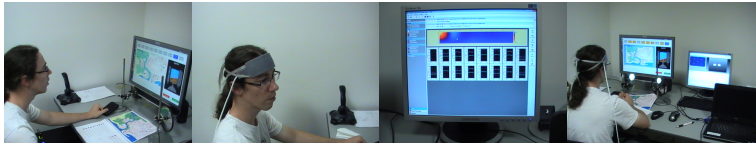
- Eye-tracker (SMI RED250)
- Participant's gaze monitoring



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

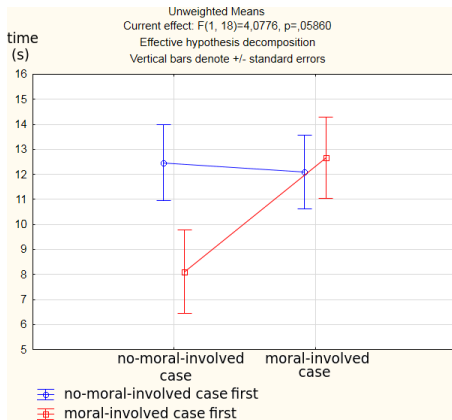
Reaction time

- 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

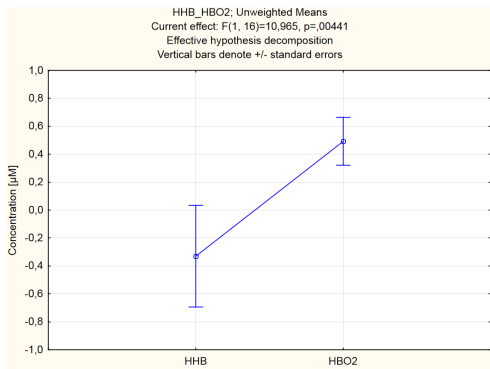
X Hyp. 3: response time is dependent from event order



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

Physiological Results

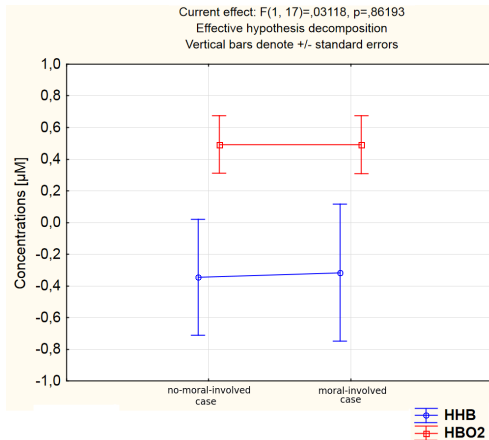
✓ Hyp. 1? (Prefrontal cortex activity may increase)



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

Physiological Results

X Hyp. 1: cannot be confirmed



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

Discussion

- X Hyp. 1: Prefrontal cortex activity may increase when facing an impersonal moral dilemma
- X Hyp. 2: Decision time may be longer when facing a moral dilemma
- X 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
 - evaluation of situation complexity
 - evaluation of amount of cognitive ressources required

Thank you for your attention. Any Question ?