

Ethical Guidelines for Persuasive Systems

Marco Guerini, Oliviero Stock
(ITC-irst, Trento, Italy)

Outline

- Overview
- Some Questions
- Proposal

Overview

- Overall question: who has the **ethical responsibility** over the behaviour of persuasive systems?
- We have to distinguish between:
 - “Nowadays” systems
 - “Future intelligent” systems

Systems development Continuum

- It is a continuum: we have degrees of autonomy



Some preliminary questions: System Autonomy

- Two different views of autonomy:
 - **Decisional autonomy:** How much the system can take decisions on its own
 - **Systems capable of evolving** (e.g. through learning): how to predict how they will eventually behave?
- Autonomy can be dangerous: e.g. Recommendation systems (such as Amazon's) that recommend pornographic contents to minors.

Decisional Autonomy

- **Weakest Solution:** stimuli-response mechanism *without any reasoning* (like a fire alarm).
 - E.g. a machine that detects strong emotions (via biometrical sensors) and simply calls the master
- **Weak solution:** capability of detecting dilemmas during planning process: delegation of the decision to the human with responsibility over its conduct.
- **Middle solution:** built-in plans of possible conducts for pre-defined situations. If necessary, submission to a human for the final decision.
- **Strong solution:** capability of making own decisions.

State of our work

- “Nowadays” systems →
ethical responsibility of developers
- “future intelligent” systems →
ethical responsibility of the system itself

- We focused on the second group: Ethical
Persuasive autonomous agents.

Outline

- Overview
- Some Questions
- Proposal

How to proceed

- Nowadays and future intelligent systems are different: guidelines for developers on ethical behaviour vs. guidelines for developers on autonomous ethical judgment.
- Yet: use analysis of “future intelligent” systems for guidelines on nowadays systems.

Some preliminary questions: Developer's Responsibility

- Can we say that the developer is “ethical” if the system behaves ethically: e.g. ethical behaviour of the machine but unethical motivations of developers.
- Problem of a group of people working each on different parts (e.g. maybe every single part is ethical but the overall behaviour of the system is not). Who is in charge for the overall ethicality?
- For “future intelligent” machines: can we imagine a sort of **ethical Turing test**?

Some preliminary questions: The scenario

- Risk of putting our systems “out in the wild”.
Just in the lab or in the real world? Ethical guidelines:
 - Just in the lab → OK
 - Out in the wild → Not enough
- ... but in the lab we could want to look exactly at unethical situations or unforeseen consequences of the guidelines adoption.

Outline

- Overview
- Some Questions
- Proposal

Ethical Guidelines: Form Proposal

- A form that covers sensitive topics regarding persuasive HCI applications.
- It is meant as an aid for researchers to **become aware** of potential issues and drawbacks of their applications.
- No indications of how to behave, but eventual involvement of ethical committee

Two-layered problem

- Ethicality of persuasion:
 - Ethicality of the communication (ax)
 - Ethicality of the induced action (ay)
- There are potential issues and drawbacks specific for the two layers.

Suggested form: ethicality of the communication

- Will your application be creatively capable of:
 - Lying? If so, why?
 - collecting info about users other than expected ones? If so, why?
 - Hiding important information? If so, why?
 - Hiding its true intentions? If so, why?
 - Overemphasising an emotional state? If so, why?
 - Inducing extreme emotions? If so, why?
 - Inducing negative emotions? If so, why?
 - Altering users' emotional state subliminally or against the users' free and conscientious choice? If so, why?

Suggested form: ethicality of the induced action

- Will your application be creatively capable of:
 - Preserve the interest of the receiver? If not, why?
 - Induce users to act against their (permissible) interests? If so, why?
 - Use its influence over the users for affecting third parties (positively or negatively)? If so, why?

Acknowledgements

- We would like to thanks Sabine Döring, Peter Goldie and their students for the useful suggestions and comments.