

## EETAS: A PROCESS FOR EXAMINING ETHICAL TRADE-OFFS IN AUTONOMOUS SYSTEMS

CATHERINE MENON and SILVIO CARTA

*University of Hertfordshire, Hatfield, England, United Kingdom*

*E-mail: c.menon@herts.ac.uk, s.carta@herts.ac.uk*

*www.herts.ac.uk*

Public-facing autonomous systems present society with significant ethical challenges, not least of which is the difficulty in conceptualizing and assessing how these systems balance competing ethical principles. We present EETAS: a structured, gamified process for obtaining input from end-users and the general public into the ethical trade-offs and balances made by a specified autonomous system. We describe how the outcomes of the EETAS process can be used to inform the design of autonomous systems, as well as improve the general public's understanding of ethical complexities.

### 1. Introduction

One of the most complex obstacles to public acceptance of autonomous systems (AS) is the lack of understanding of how these systems handle competing ethical principles. We describe ongoing work to develop a structured, gamified process for obtaining stakeholder input into the potential ethical trade-offs relevant to a given AS. This process will enable early-lifecycle collaboration between developers and stakeholders to flag these trade-offs and determine any constraints necessary to render them acceptable. The benefits of this process are two-fold: enabling the ethical perspectives of stakeholders to influence system design, and providing a gamified way for the public to understand ethical complexities of AS.

### 2. EETAS: Examining Ethical Trade-offs in Autonomous Systems

#### *Step 1: Provide AS description*

A participant group is selected, including developers of the AS under consideration, proposed end-users, regulators and members of the public. The developers provide the group with a written, accessible description of the AS and its relevant functions. Appropriate descriptions may specify, for example, that this is “an assistive robot that reminds you when to take medication, alerts you when you have left the oven on and engages you in conversation”.

#### *Step 2: Identify relevant ethical principles*

Participants are divided into teams and each team is provided with a set of pre-prepared cards, listing ethical principles generally relevant for a variety of AS. We have identified an initial set of principles from a literature review of existing and developing standards, including [1], [2] and [3]. Two sample principles from this set are:

- System promotes human physical safety
- System obeys human commands

Teams are asked to select those principles from this set they consider relevant for this specific AS, and provided with blank cards should they wish to identify additional ethical principles

#### *Step 3: scenario construction*

Teams are then asked to generate scenarios for this system in which two of the selected ethical principles are in conflict with each other. For example, a team may propose the scenario where a user commands their assistive robot not to remind them about medication, creating a

conflict between the ethical principles of “system promotes human physical safety” and “system obeys human commands”. A set of guide words based on Hazard and Operability Analysis processes will be provided to the teams to assist in brainstorming these scenarios.

#### Step 4: Identify constraints

Teams then swap scenarios, and each team works collaboratively to identify design, environmental or end-user constraints under which they would accept different balances between the ethical principles in each scenario. To gamify this step, each team is allocated points for every scenario in which they identify constraints that render at least two different balances acceptable. Teams should be asked to vote on whether they think these constraints are feasible and additional points allocated accordingly. The following prompts can be used to aid teams in identifying feasible constraints:

- “Would you accept this balance if users were made aware of it beforehand?”
- “Do you think the person benefitting in this scenario has the moral right to do so?”
- “Could this balance work in a different environment or with different users?”

#### Use of a design tool

Steps 2 – 4 are to be performed with the aid of a pre-prepared design tool, EETAS-Trade-Offs-for-You (EETAS-TOY), which represents the AS by a solid block and the relevant ethical principles as sliding bars, as in Figure 1. Participants connect bars end-to-end to represent ethical trade-offs and to discuss how different principles may be prioritized in each scenario.

#### Recording outcomes

The outcomes of each step are to be recorded using techniques such as mind-mapping. These records can then be used by the AS developers to identify further design requirements which enable or implement the constraints identified by each team. The EETAS-TOY tool itself may be retained by user organizations to aid in explaining ethical trade-offs.

### 3. Conclusions and next steps

A prototype of the EETAS-TOY tool has been built, and a preliminary pilot study of EETAS is currently being performed in order to determine the effect of this gamified approach and yield some initial results. We plan to further validate the EETAS process via workshops involving designers, end-users and AI ethics organizations including WeandAI<sup>a</sup>.



Figure 1. The EETAS-TOY gamified tool

### References

1. British Standards Institute (2016). *Guide to the ethical design and application of robots and robotic systems*. BSI 8611.
2. Institute of Electrical and Electronic Engineers (2018). *Ethically Aligned Design*, v2, [https://standards.ieee.org/wp-content/uploads/import/documents/other/ead\\_v2.pdf](https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf)
3. Leslie, D (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute, <https://doi.org/10.5281/zenodo.3240529> 4.

<sup>a</sup> [www.weandai.org](http://www.weandai.org)