Preface

Grégory Bonnet[*] Maaike Harbers[†] Koen Hindriks[‡] Mike Katell[§] Catherine Tessier[¶]

The development of Artificial Intelligence is experiencing a fruitful period of incredible progress and innovation. After decades of notable successes and disappointing failures, AI is now poised to emerge in the public sphere and completely transform human society, altering how we work, how we interact with each other and our environments, and how we perceive the world. Designers have already begun implementing AI that enables machines to learn new skills and make decisions in increasingly complex situations. The result is that these machines - also called intelligent agents - decide, act and interact in shared and dynamic environments under domain constraints, where they may interact with other agents and human beings to share tasks or execute tasks on behalf of others. Search engines, self-driving cars, electronic markets, smart homes, military technology, software for big data analysis, and care robots are just a few examples.

As intelligent agents gain increased autonomy in their functioning, human supervision by operators or users decreases. As the scope of the agents activities broadens, it is imperative to ensure that such socio-technical systems will not make irrelevant, counter-productive, or even dangerous decisions. Even if regulation and control mechanisms are designed to ensure sound and consistent behaviors at the agent, multi-agent, and human-agent level, ethical issues are likely to remain quite complex, implicating a wide variety of human values, moral questions, and ethical principles. The issue is all the more important as intelligent agents encounter new situations, evolve in open environments, interact with other agents based on different design principles, act on behalf of human beings and share common resources. To address these concerns, design approaches should envision and account for important human values, such as safety, privacy, accountability and sustainability, and designers will have to make value trade-offs and plan for moral conflicts. For instance, we may want to design self-driving cars to exhibit human-like driving behaviors, rather than precisely following road rules, so that their actions are more predictable for other road users. This may require balancing deontic rule-following, utility maximization, and risk assessment in the agent's logic to achieve the ultimate goal of road safety.

Questions to be asked here are: How should we encode moral behavior into intelligent agents? Which ethical systems should we use to design intelligent, decision-making machines? Should end-users have ultimate control over the moral character of their devices? Should an intelligent agent be permitted to take over control from a human operator? If so, under what circumstances? Should an intelligent agent trust or cooperate with another agent embedded with other ethical principles or moral values? To what extent should society hold AI researchers and designers responsible for their creations and choices?

This workshop focuses on two questions: (1) what kind of formal organizations, norms, policy models, and logical frameworks can be proposed to deal with the control of agents' autonomous behaviors in a moral way?; and (2) what does it mean to be responsible designers of intelligent agents? The workshop welcomes contributions from researchers in Artificial Intelligence, Multi-Agent Systems, Machine Learning, Case-based reasoning, Value-based argumentations, AI and Law, Ontologies, Human Computer Interaction, Ethics, Philosophy, and related fields.

The topics of interest include (but are not limited to):

- machine ethics, roboethics, machines and human dignity
- reasoning mechanisms, legal reasoning, ethical engine

[*]University of Caen Normandy, UMR CNRS 6072 GREYC, France, email: gregory.bonnet@unicaen.fr
[†]Delft University, Interactive Intelligence Group, The Netherlands, email: m.harbers@tudelft.nl
[‡]Delft University, Interactive Intelligence Group, The Netherlands, email: k.v.hindriks@tudelft.nl
[§]University of Washington, Information School, USA, email: mkatell@uw.edu
[¶]Onera, France, email: catherine.tessier@onera.fr

- authority sharing, responsibility, delegating decision making to machines
- organizations, institutions, normative systems
- computational justice, social models
- trust and reputation models
- mutual intelligibility, explanations, accountability
- consistency, conflicts management, validation
- philosophy, sociology, law
- applications, use cases
- societal concerns, responsible innovation, privacy Issues
- individual ethics, collective ethics, ethics of personalization
- value sensitive design, human values, value theory

Ten papers were submitted to EDIA, nine of them have been accepted for presentation after being reviewed by three or four members of the Program Committee. The accepted papers have been organized in two sessions:

1. Ethical issues and ethical application of intelligent agents (four papers)
2. Ethical models of intelligent agents (five papers)

The EDIA workshop would not have been possible without the support of many people. First of all we would like to thank the members of the Program Committee for providing timely and thorough reviews. We are also very grateful to all the authors who submitted papers to EDIA Workshop. We would like also to thank Bertram Malle and Jeremy Pitt who have accepted to give an invited talk in the workshop. We would like also to thank the organizers of ECAI 2016.

Program committee

- Huib Aldewereld, Delft University of Technology, The Netherlands
- Mark Alfano, Delft University of Technology, The Netherlands
- Peter Asaro, The New School, USA
- Olivier Boissier, Mines Saint-Etienne, France
- Tibor Bosse, Vrije Universiteit Amsterdam, The Netherlands
- Gauvain Bourgne, Universit Pierre et Marie Curie, France
- Selmer Bringsjord, Rensselear Polytechnic Institute, USA
- Joanna Bryson, University of Bath, UK
- Pompeu Casanovas, Royal Melbourne Institute of Technology, Melbourne, Australia
- Nigel Crook, Oxford Brookes University, UK
- Michal Dewyn, Ghent University, Belgium
- Sjur Dyrkolbotn, Durham University and Utrecht University, UK and The Netherlands
- Isabel Ferreira, University of Lisbon, Portugal
- Jean-Gabriel Ganascia, Universit Pierre et Marie Curie, France
- Pim Haselager, Radboud University, The Netherlands
- Marilena Kyriakidou, Coventry University, UK
- Bertram Malle, Brown University, USA
- Pablo Noriega, Intitut d'Investigaci en Intelligncia Artificial Barcelona, Spain
- Jeremy Pitt, Imperial College London, UK
- Thomas Powers, Center for Science, Ethics and Public Policy, USA
- Lambr Royakkers, Eindhoven University of Technology, The Netherlands
- Giovanni Sartor, European University of Florence, Italy
- Aimee van Wynsberghe, University of Twente, The Netherlands
- Pieter Vermaas, Delft University of Technology, The Netherlands

Organization committee

- Grégory Bonnet, Normandy University, France
- Maaike Harbers, Delft University of Technology, The Netherlands
- Koen V. Hindriks, Delft University of Technology, The Netherlands
- Michael A. Katell, University of Washington, USA
- Catherine Tessier, Onera, France