Society for the Advancement of Philosophy

ZAGREB APPLIED ETHICS CONFERENCE 2017

The Ethics of Robotics and Artificial Intelligence

PROGRAM & ABSTRACTS

June 5–7, 2017
Matica hrvatska – Matrix Croatica
Strossmayerov trg 4 · Zagreb · Croatia
www.upf.hr/en/zaec-2017

Monday, 5 June 2017

09:00–09:15 Opening of the conference

TOMISLAV BRACANOVIĆ, *President of the Organizing Committee*
TOMISLAV JANOVIĆ, *President of the Society for the Advancement of Philosophy*

09:15–10:15 Plenary lecture

BOJAN JERBIĆ, *University of Zagreb*
Medical robotics – step ahead of known concepts of ethics

10:15–10:35 Coffee break

10:35–12:05 Session I

FRIDERIK KLAMPFER, *University of Maribor*
Moral and policy issues in programming autonomous cars for decision-making in cases of unavoidable harm

DIJANA MAGĐINSKI, NINO TOLIĆ
The *who* question of autonomous cars

JAVIER RODRÍGUEZ-ALCÁZAR, *University of Granada*
Do driverless vehicles face moral dilemmas?

12:05–13:30 Lunch break

13:30–15:00 Session II

DAVOR PEĆNJAK, *Institute of Philosophy, Zagreb*
Responsibility and autonomous weapons systems

MIGUEL MORENO MUÑOZ, *University of Granada*
The viability of "embedded ethics" in robotic military systems without humans in the decision loop

LUKA OMLADIČ, *University of Ljubljana*
???

15:00–15:20 Coffee break

15:20–16:50 Session III

MIRKO DANIEL GARASIC, *IMT School for Advanced Studies Lucca*
The mechanization of love

AURA-ELENA SCHUSSLER, *Babeş-Bolyai University, Cluj-Napoca*
Sexbots and the issue of sexual solipsism – new ethical perspectives

PAULO ALEXANDRE E CASTRO, *University of Algarve*
Sexbots are really social robots? Some considerations on the impossibility of ethics

Tuesday, 6 June 2017

09:00–10:00 Plenary lecture

GIANMARCO VERUGGIO, *Italian National Research Council, Genoa*
???

10:00–10:20 Coffee break

10:20–11:50 Session IV

JAMES DIGIOVANNA, *City University of New York*
Artificial friends

ROSANGELA BARCARO, MARTINA MAZZOLENI, PAOLO VIRGILI,
Italian National Research Council, Genoa / Ministry of Justice, Genoa
Ethics of care and robots care givers: the evolution of robots through human care giving

ZOE PORTER, *University of York*
Eudaimonia for older people in the robotic age

11:50–13:30 Lunch break

13:30–15:00 Session

TOMISLAV JANOVIĆ, *University of Zagreb*
What (if anything) can we learn from mindreading robots?

MAIKE KLEIN, *University of Stuttgart*
The moral relevance of emotions in artificial systems

FABIO FOSSA, *University of Pisa*
Between functioning and acting. "Autonomy" and "morality" in human beings and machines

15:00–15:20 Coffee break

15:20–16:50 Session

IHSAN BARIS GEDIZLIOGLU, *John Cabot University, Rome*
What they can do & what they should do?

LILIAN BERMEJO-LUQUE, *University of Granada*
The only rule that a super intelligent robot must obey

CANSU CANCA
RECs for artificial intelligence: an unintelligent choice

18:30 Guided sightseeing of Zagreb for participants of the conference

Wednesday, 7 June 2017

09:00–10:30 Session

ANDREAS MATTHIAS, *Lingnan University, Hong Kong*
Moral imperialism and the social dynamics of human-robot interactions

FRANCISCO LARA SÁNCHEZ, *University of Granada*
Artificial intelligence and moral enhancement

TOMISLAV MILETIĆ, *University of Rijeka*
Moral enhancement and distributed intelligence: Is the age of Plato's guardians upon us?

10:30–10:50 Coffee break

10:50–12:20 Session VII

ANÍBAL MONASTERIO ASTOBIZA, *University of Oxford*
The ethics of AI and AI based modelling of ethics

SANDRO SKANSI, *University College Algebra, Zagreb*
Deep learning and the rise of connectionism in artificial intelligence

KRISTINA ŠEKRST, *University of Zagreb*
Nazi sex robots: moral reasoning guided by computational complexities

12:20–13:50 Lunch break

13:50–15:20 Session VIII

MARCIN GARBOWSKI, *John Paul II Catholic University of Lublin*
The personalist proposal for a solution for AI ethics dilemmas

JUDITH ZINSMEIER, *University of Tűbingen*
Human-computer interaction – A critical consideration with recourse to Hannah Arendt's concept of action

ROSALLIA DOMINGO, *Central European University, Budapest*
Cyborg embodiment: Exploring the concept of body and technology hybridity in Mamoru Oshii's *Ghost in the Shell*

15:20–15:40 Coffee break

15:40–16:40 Session

MARTIN GLICK, *University of Göttingen*
The ethical repercussions of a spectrum of programming power

JASMIN POLJAK, JOSIP NAGLIĆ, *Natural Science and Graphics School Rijeka*
The division of responsibility between the publisher and the user of upgrade

17:40 Closing of the conference

20:30 Conference dinner

ABSTRACTS

Sexbots are really social robots? Some considerations on the impossibility of ethics

PAULO ALEXANDRE E CASTRO
University of Algarve

Human reality is constantly changing, and his sex life as well. As Anthony Giddens, N. Luhman and so many others saw, intimacy changed since the XVIII century, in every aspects of human life (familiar, social and economic). Now, we may think that we are experiencing a new and probably last transformation of intimacy with the introduction of sexual robots. They are not scientific fiction anymore (as in movies) but they will become a reality in a few decades. Such a reality, as some of the experts say, may change the scenario of humanity since they are presented as perfect, always ready and non boring lovers. Faced with this new reality, the abolition of sexuality as a phenomenon and experience of love (human) will certainly be modified, or can we just talk about a new kind of love or just a new way of seeing the arriving of a new sexuality? Since they can replace easily human interaction, two major questions arise: it is still possible to talk about intimacy? And, is it possible to develop an ethics for sexbots? Our essay will try to answer these questions.

Ethics of care and robots care givers: the evolution of robots through human care

ROSANGELA BARCARO, MARTINA MAZZOLENI, PAOLO VIRGILI
Italian National Research Council, Genoa / Ministry of Justice, Genoa

There's been recently a strong upheaval of interest about artificial intelligence (A.I.) and its possible developments. Many A.I. applications are already in use in many fields of our daily life: research engines, domestic robots, deep see exploration vehicles, healthcare devices. The most amazing between them are not medical equipment and devices, that can help physicians improve their diagnostic skills, but the development of particular robots used for taking care of chronically ill patients and/or elderly people. The use of state-of-the-art technology has normally determined an improvement of healthcare quality both for patients, physicians and healthcare givers. Until today, human beings have been applying technological instruments and devices in what we could define as a "passive" aided mode: artifacts

being used by human subjects. We are currently switching to an "active" aided mode, where technological aids are no longer simply used by humans, but are designed to act autonomously. That opens an entirely new area of interest and analysis: we are no longer facing ethical problems related to single patients and their therapies and care, but a new whole class of problems related to couples patient-artifact, where the human moral agent and the artifact are symbiotic, and the distinction between them is somehow blurred.

The only rule that a super intelligent robot must obey

LILIAN BERMEJO-LUQUE
University of Granada

The idea that, in the future, a robot far more intelligent than us could take decisions that affect us has raised all kinds of fears. Initiatives – from Asimov's famous *Three Rules for Robotics* (1942) to the *Asilomar AI Principles* (2017) – have been launched in order to ensure that the development of AGI is beneficial, or at least safe, for humans. Yet, as Bostrom (2014) among others has urged, we must be very careful with the wishes that we implement in this respect, because as it happened to King Midas, any such wishes may backfire. In this paper I analyze current attempts at making the development of AGI safe, present their limitations and propose the following alternative: securing that no AGI agent takes novel decisions without argumentatively persuading those to be affected. Apart from being simple and relatively easy to implement, this proposal has important advantages: it avoids backfires while being flexible enough to incorporate the possibility that future generations (namely, those coexisting with AGI technology) have values substantively different from ours. The latter feature is important because we don't know how long will it take to have AGI devices, but we must take action in matters of security well in advance this technology is available; yet, it is foreseeable that a society with such a revolutionary technology changes its values and preferences in unexpected ways. In addition to these advantages, the argumentation rule would reinforce the dignity of both humans and super-intelligent creatures and the mutual benefit in their interactions.

RECs for artificial intelligence: an unintelligent choice

CANSU CANCA

With the developments in artificial intelligence (AI), the need for a robust ethical framework for AI research becomes pressing. AI research presents unconventional ethical questions, that can be grouped into two main categories: (1) moral

justification of AI behavior, and (2) moral status of AI. The first category is concerned with ethical judgements that are incorporated in designing a wide range of AI systems. Most notably, the role of ethical judgements is crucial in 'autonomous' AI systems that are also tasked to make moral decisions. Yet, even in designing 'simple' AI systems, there is room for incorporation of ethical judgements (as well as biases). Validity of these ethical judgements affect the ethical status of the resulting AI actions. The second category of AI specific ethical questions, on the other hand, focuses on AI's capacities as it develops further, and the resulting moral status of AI. If and when the AI acquires animal-level or human-level moral status, our obligations towards AI would change presenting new limits on moral permissibility of AI research. To ensure that AI research conforms to ethical principles, current draft guidelines suggest implementing a system for AI research ethics governance that is similar to human subject research: oversight by research ethics committees (RECs). However, these suggestions not only overlook deep normative and practical problems in the current system of research ethics governance but also fail to address its inadequacy in handling AI specific moral issues. Specifically, this paper argues against the ethics oversight system, and instead proposes a collaborative system among scientists and ethicists.

Artificial Friends

JAMES DiGIOVANNA
City University of New York

While robots were originally conceived of as laborers, advances in AI and emotional modeling have led to caregiver and "companion" robots. The companion-robot is meant not only as a caregiver, but also to replace some of the interactions one would have had with family and friends in more communitarian societies. However, robot-friends, unlike organic ones, are quickly reprogrammable. If a real friend wants to do something we don't want to do, or holds an opinion contrary to ours, or even shows less-than-desired interest in our grandchildren, we cannot simply throw a switch to make them more agreeable. But with a robot's reprogrammability, that switch is there, waiting for us to "fix" our friend into someone more in keeping with our desires. With human friends, we learn to compromise and negotiate, and we expand our tolerances and interests in order to maintain the friendship. By eliminating this work of friendship, the robot is in some ways an ideal friend. This idealization, though comes with the loss of important moral elements of the friendship: it's by attuning ourselves to our friends, compromising, and respecting the autonomy and otherness of the friend that we grow emotionally and morally in the friendship. Artificial friends could replace this growth with a video-game-like instant gratification. In addition to the moral and emotional losses, losses in character virtues like flexibility and spontaneity, and epistemic virtues like the ability to see

from another perspective, could be harmed, leaving us lacking many of the essential social traits that constitute personhood.

Exploring the concept of body and technology hybridity in Mamoru Oshii's *Ghost in the Shell*

ROSALLIA DOMINGO
Central European University, Budapest

More than thirty years after the publication of Donna Haraway's *Cyborg Manifesto*, the image of cyborg still strikes as an interesting way to imagine the self in a postmodern context. The cyborg figure has continued to inform feminist reading and articulation of the mind/body bifurcation in science fiction film and literature that epitomizes the body-as-machine metaphor. The embodied hybridity of Haraway's cyborg particularly opens up a move beyond the dualistic epistemologies that produced antithetical subject positions that feminist articulations of science fiction take up in highlighting the paradoxical representations of embodiment and subjectivity in cyborg technology. Inasmuch as contemporary cyborg theory reaffirms the basic foundations of Cartesian dualism in late twentieth century cyborg films, Mamoru Oshii's film *Ghost in the Shell*, as a reference to the philosophical concept "ghost in the machine" could be a useful site of interrogating the body and technology hybridity as a place where we project a normativization of the body. It is in this context that I would like to explore the female, posthuman body presented in the film as texts coded with cultural and gendered anxieties. I aim to look into how the film, through its depiction of the cyborg heroine, Major Motoko, implicates female bodies inscribed with cultural and social rhetoric. To this end, I will use Lauren Wilcox's perspective of technology as both embodied and embodying. I shall focus on Wilcox's use of the posthuman theories by Donna Haraway and Katherine Hayles in understanding materialization/dematerialization of technology in the film.

Between functioning and acting. "Autonomy" and "morality" in human beings and machines

FABIO FOSSA
University of Pisa

The complex interrelation between humans and machines is usually framed in terms of similarity. But how far can this be taken? Some have suggested that, since machines are technological reproductions of human modes of existence, the theoretical models on which they are based shall apply also in the human case. If this

were so, it would be more appropriate to frame the relationship in terms of identity rather than of similarity. From a moral perspective, this means that machines and human beings should be considered as instances of a same kind, that of 'moral agents', which would differ only by degree of completeness. In my talk I will try and show why this claim is mistaken. In order to do so, I will draw on two different conceptual structures which can account for what we perceive as "representationally based behaviour" (i.e., self-regulating or "teleological" behaviour). I will first introduce the concept of 'function', as well as the main features of that of 'action'. I will then pinpoint the different ways the acting or functioning entity «*has* purpose and *acts on* purpose», as Hans Jonas suggested. I will argue that, whilst human modes of existence fall under the concept of action, what machines do should be strictly interpreted by reference to the concept of function. I will conclude by opposing the notion of Artificial Moral Agent (AMA) and suggesting how, in my opinion, the question regarding the moral status of machines should be phrased.

The mechanization of love

MIRKO DANIEL GARASIC
IMT School for Advanced Studies Lucca

In a series of articles, Brian Earp, Anders Sandberg and Julian Savulescu have identified a very specific ramification of the possible applications of enhancing biotechnologies: love drugs. I define this kind of enhancement as Emotional Enhancement (EE). EE has a number of peculiarities: not last, the fact that – differently from most forms of enhancement – more is not necessarily better. In line with this assertion, it has already been accepted by the same authors that diminishment might be in fact the answer to our problems in certain instances. Ultimately, EE wants to ensure the overall well-being of the individual, and thus – although deeply consequentialist in its core – the authors claim it to be not obsessed with absolute numbers per se. What counts is the impact that such biotechnologies can have in the overall life experience of the adult subject freely and competently choosing to undergo the procedure in question. I take issue with this way of conceptualizing and framing the impact that EE can have on our society through its entrance in our love relationships, and I propose introducing in the debate the definition of "mechanization of love" as a useful term to understand more clearly the dynamics in place.

The personalist proposal for a solution for AI ethics dilemmas

MARCIN GARBOWSKI
John Paul II Catholic University of Lublin

Many critics of transhumanist and of radical technological enhancement movements express fears that new entities (robots, and especially various forms of emerging sentient "digital life") may become dangerous to humans. In response advocates of those movements suggest installing into new entities an axiom, akin to Isaac Asimov's laws of robotics – "robots may not harm a human person". In my paper I attempt to show that it is not enough. With the status of what is a human being left to the discretion of the law, just as it is done in most Western societies, the consequences may prove catastrophic for humanity. If an agent deciding on whether an entity is worthy of the status of a "human being", "person" or "citizen" is to be a pre-programmed automaton, it might decide to choose other criteria which are easier to turn into objective traits than the arbitrary ones typical for positive law. The solution might be to use the axiom of the inviolability of human life as such based in the concept of the dignity of human being as developed by philosophical personalism.

What they can do & what they should do?

IHSAN BARIS GEDIZLIOGLU
John Cabot University, Rome

"Can a machine feel?" Everyone would ask this question, and rightly so. For, we generally attach a moral motive behind feelings. Some has claimed that it was feeling and empathy that made us moral animals. In light of this, consider. Neurologists have postulated a simple and efficient way to understand if a given individual can be identified as a sociopath (lacking the ability to empathy) or not. To a stimulus of suffering of a third party, everyone would say, "I am sorry" or alike. If an individual, says while only the cerebral cortex (language centre) in his brain is used and not the limbic system (emotional centre), ergo, he is sociopath. The underlying idea in this is that a sociopath, even though does not empathise with the suffering, knows how to react to it. In such cases, they are called 'functional sociopaths'. This distinction is crucial, as we advocate, to the AI research. For, a machine would not be able to experience pain as we do (cf. Wittgenstein's beetle in a box argument), yet, the only way we can consider a machine to be functional is under the condition that it acts *as if* it were feeling. In other words, we can return to the first functionalist premiss and claim that as long as the same output is produced for a given input in different systems, i.e. as long as a machine can act as if it has feelings, we have to consider it functional at least as a sociopath.

The ethical repercussions of a spectrum of programming power

MARTIN GLICK
University of Göttingen

Pioneering work by Ugo Pagallo, Peter M. Asaro, but especially John Danaher with his views on Retribution highlights varying degrees of liability involved in robotics. But it is ownership entitlement on the part of the consumer and initial programming on the part of the manufacturer that have not been given a fair account. Contemporary uses of robotics, specifically in the field of well-being and care-taking focus on what I would like to call "augmented activity" which depends on a relation between the three parties in their shared interactions with the robotic device. From the most basic wholly manufacturer-guided (normative) imperatives to the kind where consumers are allowed to input their own (practical) wishes, finally ending in the autonomous robot's (self-sufficient) activity, each party is responsible for varying spheres of imperative-making decisions and these come with varying degrees of moral responsibility. Between the three parties involved in the production and use of household robots there is an asymmetrical and descending relation concerning the kind of software changes or programming power that each has. I outlined that the manufacturer has the totality of software programming available to them which guide a robot's general normative imperatives, upon which the user or consumer can make very few software changes based upon practical or realworld wishes. Finally the robot itself will, at least in the near future, be granted the least amount of control over its software. The few pieces available concerning robot Agency and Liability highlight the use of robots in the field but simplify the scenario heavily.

What (if anything) can we learn from mindreading robots?

TOMISLAV JANOVIĆ
University of Zagreb

One of the more exciting developments in contemporary robotics and AI is the construction of robots sensitive to internal states of others (including false beliefs) and programmed to act upon these states. There are two (very different) kinds of cues that such devices can make use of: neurological (electrical activity in specific brain regions) and behavioral (body signals such as facial expression, direction of gaze, pitch of voice, etc.). Despite their apparent naivety, these attempts entice speculations about a new era of human-robot interaction. The prospect of mass production of artificial agents capable of representing/interpreting internal states (equivalent to emotions, desires, beliefs, intentions, etc.) – both their own and others' – and making use of this capacity to predict, evaluate and direct actions, deserves a philosophical scrutiny. In my contribution, instead of trying to envision all the

interesting possibilities and challenges opened by the development of such a technology, I want to focus on some current attempts at constructing mindreading robots and examine whether they should have any bearing on (1) our actual understanding of minds and (2) our notions of autonomy and moral agency.

Medical robotics – step ahead of known concepts of ethics

BOJAN JERBIĆ
University of Zagreb

Robots are good, despite the general perception that robots are stealing our jobs or they are dangerous for our humanity, which is mostly influenced by science fiction and the common fear of new coming technology. Generally, robots improve our quality of life, reduce production labor and costs, improve product quality and working conditions and reduce the waste of resources. However, the robotics in the context of artificial intelligence opens up new technological and even cultural concepts known as autonomous cognitive machines. The autonomous machines are artifacts that are able to make independent decisions and shape their behavior. This is a completely novel situation in the history of mankind, out of our common experience and cultural heritage. For the first time, we are supposed to deal with own products able to act without our control. The anthropomorphism of robotic technology combined with cognitive abilities additionally emphasizes and complicates the assimilation of robotic technology in our daily lives. The impact goes from economic, social to cultural aspects. Therefore, new approaches and new understanding of science and culture need to be adopted: changing from "data driven" to "AI things driven" concepts. The rise of intelligent robots implies more human activities in their reach, the activities we never believed could be performed by machines. The healthcare is typical example, because it is the most prominent form of human behavior. The use of robots in medicine seems like a huge potential for improving various technically and/or physically demanding medical procedures and specific skills that doctors must possess in addition to theoretical and experiential knowledge in the scope of their profession. However, the use of robots in medicine, despite numerous tangible benefits, is faced with numerous scientific, technological and ethical challenges. The relationship of humans toward artifacts (material goods) has always been a socially regulated. Although, the relationship of machines/artifacts to people seems more significant now when dealing with autonomous robots which behavior is based only on our expectations. Can such intelligent robots develop their own ethical standards? Can a robot understand culture? There are many questions, but answers are few. The lecture will highlight the main directions of the development of surgical robotics, possible advantages of application as well as problems. Special attention will be given to the Croatian project RONNA – robotic system for stereotactic neurosurgical operations, as well as its clinical application.

The discourse will be expanded through some additional development directions in the bionics and will address related ethical and moral challenges from the engineering point of view. Building an early awareness of the consequent ethical, legal, economic and social issues, will enable the easier development of our new "robotized" society and culture.

Moral and policy issues in programming autonomous cars for decision-making in cases of unavoidable harm

FRIDERIK KLAMPFER
University of Maribor

Besides significantly reducing the need for human workforce in transportation, and depriving millions of professional truck and taxi drivers of their current employment, autonomous vehicles (AVs) are also promising to increase traffic efficiency, reduce pollution, and avoid most, if not all, traffic accidents. Still, not all future crashes will be eliminated. From time to time, AVs will be faced with the kind of choices that were made prominent in the past four decades by moral philosophers: whether to run over several pedestrians or instead swerve and sacrifice a single bystander, or whether to sacrifice their own passenger(s) to save one or more other people. As unlikely as such Trolley-like scenarios may appear at this point, they will become sufficiently common with millions of AVs on the road to call for clear moral guidance. Accordingly, the algorithms in control of AVs' behaviour will need to embed moral principles that govern the distribution of unavoidable harm in these and similar situations. What complicates the choice of decision-rules for distributing unavoidable harm in future autonomous car accidents and sets them apart from the variants of the familiar Trolley case is its wider social and economic/financial aspects: the lawmakers will need to find a reasonable compromise such that it's (i) morally defensible, (ii) generates sufficient public support, and (iii) doesn't adversely affect the sale of such vehicles, or make the option of purchasing such a vehicle too unattractive to consumers. Some recent empirical research suggests that this could turn out to be a rather challenging task – for people apparently want to see other people buying and driving utilitarian, i.e. harm-minimizing vehicles, but would themselves prefer to purchase and use non-utilitarian, i.e. passengers-prioritizing vehicles. In the paper, I try to show this particular version of Prisoner's Dilemma resolvable, in principle, as long as we're willing to make reasonable compromises regarding consumers' full autonomy.

The moral relevance of emotions in artificial aystems

MAIKE KLEIN
University of Stuttgart

Emotions seem to make human behavior less predictable or controllable. At the same time, some philosophers think that emotions influence morality and human moral decision making (Hume 1739, Prinz 2007). This dilemma is something we mostly deal with implicitly on an everyday basis. But what about artificial systems? If it is true that emotions play a vital part in our day-to-day ethical life, they may be a crucial aspect of creating a fully ethical or full-fledged moral agent (Moor 2006; Wallach and Allen 2010; Misselhorn 2013). In our increasingly mechanized world where decisions are more and more complex, artificial moral agents can for instance support or model human moral decision making. This can lead to well-considered moral decisions, better performing artificial agents, and a better understanding of human morality itself. Both top-down and bottom-up-approaches face some theoretical problems in constructing artificial moral agents. In my talk, I will argue that modeling emotions into artificial systems may present a workaround to many of the problems these approaches face. This leads to ethical questions such as: "Which emotions or which parts of affectivity should we include in an artificial system?" The answers to this kind of questions highly depend on the artificial system we are dealing with and the goal it has been programmed for. This concerns, however, not only fully lifelike artificial agents we know today from science fiction, but equally more abstract moral decision making systems that may be part of e. g. autonomous cars or medical diagnostic systems.

Artificial intelligence and moral enhancement

FRANCISCO LARA SÁNCHEZ
University of Granada

Human societies have always claimed to make their citizens more moral, many of whom have also sought to do so on their own initiative. Methods have traditionally been used to improve morality such as the (dis) incentive of certain behaviors, propaganda, religious messages, ethical reflection or, above all, education. However, the results are not so encouraging, in part because these traditional methods of moral enhancement are often slow, long-term and often ineffective. This is due to serious human limitations such as poor cognitive and deliberative capacities, weakness of will or the excessive importance given to intuitions based on emotions, favoritism and prejudice. The aforementioned difficulties of the traditional methods of moral enhancement would not be so worrying if we did not find ourselves in a world, globalized and technologically unstoppable, which changes so rapidly and for which

our current morality, so limited and narrow, does not work. However, this new world can also offer us a solution to the challenges that it poses to our morality. This solution could found in artificial intelligence. The purpose of my talk will be to present a proposal on how artificial intelligence can achieve, more rapidly and successfully than traditional methods, the mentioned universal claim to make individuals more moral. To do this, I will explain what I understand by moral enhancement, what role the agent should play in the possible actions for enhancement and what functions of artificial intelligence could serve to modify human morality acceptably. I will also try to respond to possible objections to the resulting morality, especially the objection that insists on the moral importance of freedom.

The *who* question of autonomous cars

DIJANA MAGĐINSKI, NINO TOLIĆ

Rapid development of autonomous cars has brought up several ethical issues. In this paper, we will focus on the so-called tunnel problem. The tunnel problem is a special case of a trolley problem, in which an agent must decide how an autonomous car should react in situations where fatalities are unavoidable. As in the general trolley case, a question can be raised about the autonomous car's reaction. Should the car prefer the life of a driver over a pedestrian? However, in the tunnel problem case, there is an additional question about who should decide how the car will react. Should the decision be left to the manufacturer/policymakers or should it be left to the car's owner/driver? We will first briefly examine arguments for both options as well as potential problems. After that, we will look more closely at the argument proposed by Gogoll and Müller (2016). If the driver has the option to choose between a selfish setting ("kill pedestrian") and a moral setting ("kill me"), Gogoll and Müller argue that, since this situation inevitably leads to the prisoner's dilemma, drivers have great incentives to choose the selfish setting. Instead, they propose we should apply a mandatory ethics setting that aims at maximizing benefit for everyone. However, we believe that mandatory ethics setting is not a stable strategy as it is highly susceptible to defectors and therefore also leads to the prisoner's dilemma. Finally, we will propose a way to overcome the incentive to choose selfishly in the case of a personal ethics setting.

Moral imperialism and the social dynamics of human-robot interactions

ANDREAS MATTHIAS
Lingnan University, Hong Kong

The current discussion in robot ethics is confined to a very narrow subset of the problems that are occurring in human/robot interactions. In particular, consideration of the effects of the interactions between robots and humans *on the human operator* is almost entirely absent. In this paper, we take a closer look at common interaction patterns between humans and autonomous artefacts (in autonomous driving, eldercare, childcare, and war robots), and analyse them with a view towards understanding the hierarchical relations between the roles of human operator, machine, the machine's manufacturer, and democratic lawgiving and law enforcement institutions. Our analysis shows that there is considerable reason to be concerned about how interactions between individuals and machines will impact human freedom, autonomy, and dignity. These effects must be moved into the center of attention of robot ethics researchers in order to prevent causing unanticipated harm not only to individual humans, but to our understanding of what constitutes a human being and its essential freedom and dignity.

Moral enhancement and distributed intelligence: Is the age of Plato's guardians upon us?

TOMISLAV MILETIĆ
University of Rijeka

Human history is witness to a struggle of achieving cultural progress through intellectual and moral excellence. In this regard, the discussion on "Moral Enhancement" (ME) has recently sparked a fruitful academic debate. Still, the importance of technological artefacts, especially artificial systems and intelligent agents (IA) has been termed as "conventional", focusing the discussion in the direction of possible genetic and neuro-pharmaceutical investments instead. We challenge this notion by claiming that the exploration of enhancing human cognitive and moral capacities should be first achieved through a specific type of Human-AI relation. Precisely, an extended or distributed (moral) cognitive system as the most fruitful and least contested option to enrich and enhance human (moral) decision making and acting. To illustrate a practical possibility we portray the IA as an Ambient Intelligence (AmI) capable of affective computing, developing with the human agent since early childhood (age of reason). We look at the requirements (privacy, transparency, security, identity) for such a distributed cognitive system, and imagine a future scenario in which humanity is utilizing the assistance of such

IA inside a distributed and highly personalized digital network. We also consider important consequences such a network would have for moral and large scale political decisions and assert the claim that a distributed network of human AI agents – one that enhances the capacities for creative participation, critical thought, personal choice and self-determination – is the best measure against "algorithmic" or political control we could come to experience in the coming digital age.

The ethics of AI and AI based modelling of ethics

ANÍBAL MONASTERIO ASTOBIZA
University of Oxford

AI is transforming our lives for the better. It is used for image recognition, analysis in social media sites, recommendation systems in online musical or movie platforms, robotics etc. Despite all its benefits it also has a dark side. The possibility of creating intelligent machines raises many important ethical issues (Bostrom and Yudkowsky 2014) As AI approaches humans in its intelligence it must operate safely and friendly. In the first part of the talk we will present briefly some themes related to the ethics of AI and how to build the necessary ethical controls. In the second part, we will deal with the use of AI for the enhancement of moral decision making and its applications in health care. Enhancing human intelligence is a very complex and difficult task. One way to achieve it is one of merging human intelligence with artificial intelligence, the so called, symbiosis HI + AI. The singularity, the hypothesis that the development of AI will create an explosion of increasingly intelligent machines leading to the displacement of human beings, although plausible is certainly improbable. However, the collaboration and cooperation between human beings and machines is no longer science-fiction and might be the most promising way to enhance our cognitive functions, including our morality. Symbiosis HI + AI is a model in which a group of humans work effectively with a group of machines to answer relevant questions like for example, in health, drug discovery for the treatment of diseases and, in ethics, the enhancement of moral decision making. As examples of symbiosis HI + AI we would like to show the recent developments in AI applied to drug discovery and a proof of concept idea of human-machine symbiotic combination for the enhancement of moral decision-making.

The viability of "embedded ethics" in robotic military systems without humans in the decision loop

MIGUEL MORENO MUÑOZ
University of Granada

The social regulation of robotic systems with some elements of inbuilt artificial intelligence, self-propelled by land, sea or air, and capable of interacting − at least temporarily − with the physical world without human control, poses challenges of extraordinary complexity. In particular, when their characteristics make them suitable for being used in military operations as autonomous devices under specific conditions. My purpose is to do a case-study research about the viability of some elements of "embedded ethics" in different devices, with built-in sensors and a variable range of functionality, starting with Autonomous Weapons Systems (AWS). Example #1: an intelligent mine that can be activated according to the information provided by specific sensors, in order to discriminate between targets or to interact with the environment in a way that does not necessarily involve self-destruction. Example #2: advanced missiles designed to act cooperatively in the approach to the programmed target, able to detect variations in the target's position and to elude threats in real time by sensors connected to the missile guidance systems. Based on the revision of recent literature and prototypes, the expected results should give a clearer perspective about the viability of "embedded ethics" instructions in the programming of intelligent robotic systems, including those intended for military use. As a preliminary conclusion, the heterogeneity of designs, lethal capacity and degrees of functional complexity in operational contexts – highly unpredictable – reinforces the importance of preserving human intervention in the decision loop, when the lapse for the sequence of decisions makes it possible.

Responsibility and autonomous weapons systems

DAVOR PEĆNJAK
Institute of Philosophy, Zagreb

Automatization and computerization are at its leading edge concerning military systems. They tend to be more and more sophisticated and one of the main goals is to produce highly autonomus or even fully autonomus systems, for both combat and non-combat military use, and, in a longer term, for each level: tactical, operational and strategical. So, highly or fully autonomous systems will be implemented at all levels and for almost all possible tasks in military use. Of course, though today there are already some higly autonomous systems in use, and even a few fully autonomous systems, their extensive use lies in the future. But, already now we can pose the questions of responsibility concerning such systems. There is some limited but

significant experience with semi-autonomous systems and responsibility for various issues concerning them. We can extrapolate from this and reason about the responsibility about fully autonomous military systems, especially autonomous weapons systems. But, it is not a single question. First, we can reason about agreement to develop and use such systems; we can ask questions if such systems are put to use, how should they be deployed, used and stored. One of the questions may be who would be responsible for malfunction of such a system. I shall argue, overall, that these question can be answered not much differently from the questions of resposibility in other, more conventional, military matters and situations.

The division of responsibility between the publisher and the user of upgrade

JASMIN POLJAK, JOSIP NAGLIĆ
Natural Science and Graphics School Rijeka

If a robot causes significant damage to a property or inflicts bodily harm due to an update, does the responsibility lies with the publisher of the software update's code or the user of that update? The authors illustrate on the example of *Deus Ex: Human Revolution* and similar popular media that such scenario leads to pseudo antinomy. Furthermore, they show that user usually accepts update in good faith, without adequate knowledge of the update's content (*bona fide* argument), even if it is unintentional, such as most of today's updates. The user is generally conditioned to accept update of the software in order to continue using it. Today the error in update can cause a virtual damage, but in the foreseeable future it can threaten property or human life. The authors opt for more transparent update process.

Eudaimonia for older people in the robotic age

ZOE PORTER
University of York

In my paper, I will argue that the only moral justification for elderly care robots is the flourishing, or *eudaimonia*, of the older individual within a human-based care system. This, and nothing else, should inform the purpose and design of such robots, whether they are service-type robots or companion-type robots. The flourishing of the older person may be the direct outcome of robotic support (e.g. a more independent life due to physical assistance in their own home), or may be an indirect outcome (e.g. more convivial human care as a result of relieving care workers of heavy and/or unpleasant tasks). Moreover, ensuring these beneficial outcomes will

involve a refined process of balancing potential positive and negative effects (e.g. reminders about safety risks could, if taken to extreme, entail a loss of autonomy and control). In addition, even the best-designed robot can be put to ill-use, which points to the urgency of developing a robust legal framework before the robots are introduced to market. Finally, I will argue that humanoid companion robots cannot be morally justified, since they will require or cause the individual to suspend their reason, which undermines their flourishing, and that therefore this kind of companion-type robot has no place within the context of care for older people.

Do driverless vehicles face moral dilemmas?

JAVIER RODRÍGUEZ-ALCÁZAR
University of Granada

Although driverless vehicles are expected to reduce the number of casualties caused by traffic accidents, their likely implementation poses several moral concerns. One of them is: how should vehicles behave in emergency situations where human lives are endangered? Attempts to answer this question usually take for granted the analogy with moral dilemmas faced by human drivers, and the task is often formulated in terms of *teaching* vehicles to behave morally. I claim that this analogy is misleading: driverless vehicles at the crash scene make no decision, and no moral dilemma is involved. A deliberation is indeed required, but this is not a moral, but a political deliberation undertaken by lawmakers well in advance. Thinking otherwise would constitute an instance of what Bernard Williams criticized as "political moralism". Talking in terms of moral dilemmas faces the problem that there is no universal agreement concerning moral codes of conduct, so it is not clear at all what principles vehicles ought to obey. Experimental programmes have been designed to find out how different people would behave in crash contexts, but these empirical findings don't guarantee an ethically correct answer. I develop an alternative proposal based on my own conception of the relationship between ethics and politics as a "reciprocal containment". According to it, moral considerations enter only as one of the factors that lawmakers ought to take into account. Information provided by experimental philosophy can be useful for politicians, but not in the way the designers of these experiments intend.

Sexbots and the issue of sexual solipsism – new ethical perspectives

AURA-ELENA SCHUSSLER
Babeş-Bolyai University, Cluj-Napoca

The emergence of technology and pornography as part of the individual's space of incidence also affects that person's private life, raising new issues with regard to ethics and roboethics. The general objective of the study is to analyze the possible risks/benefits that might emerge in the field of ethics which the features of pornography and sexbots raises, through the issues of sexual solipsism, in which paradigm sexbots would replace human nature. According to the sexual solipsism argument (Rae Langton), pornography treats humans as objects, and objects as sexual partners. The study seeks to apply this theory of sexual solipsism to the technological paradigm which involves the existence of sexbots. Thus, through a substitution of species, a new sexual solipsism hypothesis emerges, where pornography treats humans as sexbots, and sexbots as sexual partners (TrueCompanion's robots case study). Putting aside the theory where humans are treated as objects, the study raises the following question – are sexbots objects which pornography urges us to treat as sexual partners? If the answer is yes, then the issue of Langtonian sexual solipsism doesn't suffer major changes. However, it raises a (robo)ethical question – if we do, or do not, have the right to treat sexbots as objects? If the answer to the original question is no, it results in a major shift in paradigm at the ethical level, which pornography opens, granting sexbots the possibility of treating humans as objects. This leads to another question – what is the ethical impact/risk of such a situation on human nature?

Deep learning and the rise of connectionism in artificial intelligence

SANDRO SKANSI
University College Algebra, Zagreb

From its inception at the Dartmouth conference up to 2006 when Deep Belief Networks (Hinton et al.) were first presented, the symbolic/logical tradition in artificial intelligence was considered to be the main approach to artificial general intelligence. During this whole period, connectionism was present, but was seen mainly as a signal processing technique useful for sensors, but useless for higher functions. Recent advances in connections models, after 2006 made ever bolder claims of achieving general intelligence. The most recent advances from 2016 (Weston et al.) show that a new neural architecture is able to achieve remarkable accuracy on a set of natural language tasks, from summarization to reasoning, which puts neural models at the core of general AI today. We will offer a preliminary report on our research which builds upon the Weston et al. 2016 paper and the bAbI dataset

presented there, and propose a couple of additional tasks modelled after the bAbI dataset to address moral reasoning.

Nazi sex robots: moral reasoning guided by computational complexities

KRISTINA ŠEKRST
University of Zagreb

Computational complexity classifies computer science problems according to their inherent difficulty, i.e. resources needed to come to a solution. Well-known cases are P and NP problems, former are solvable in polynomial time, while the latter are only verifiable in polynomial time, since the time required increases at least exponentially with the size of the problem. In the field of artificial intelligence, the most difficult problems are AI-complete, and they reflect a similar status as NP-complete problems, being difficult or probably impossible to solve by standard algorithms. AI-complete problems include computer vision-related issues, natural language understanding, and similar real-life reasoning. Could an artificial intelligent agent ever develop a sense of ethics and act accordingly? Current AI systems encounter the issue of lacking common-sense knowledge of the situation, while human beings have background experiences to recognize unusual situations and adjust with ease. Ethical conundrums and different categories of various issues, being presented as real-life situations or as a problem for natural language processing, seem to belong to AI-complete problems. Thus, their resolution depends on the resolvement of the greatest computational issue: can NP-complete problems be reduced to P problems? The modern techniques of neural networks and deep learning can only give a limited notion of understanding, which is often prone to corpus-related problems, again derived from human experiences. Hence, if it is shown that higher-rank problems cannot be reduced to quickly solvable ones, even a marginal grasp of an ethical issue in question still cannot be possible by an artificial intelligent agent.

Human-computer interaction – A critical consideration with recourse to Hannah Arendt's concept of action

JUDITH ZINSMAIER
University of Tübingen

The term "human-computer interaction" is in the meantime established in the technical sciences as well as in non-technical academic disciplines. As many other terms in AI research – for instance "knowledge", "planning" and "action" – the term

“interaction” originates from the human realm and has been applied to the technical realm. This application is supposed to mirror the development of technology which is not simply a passive instrument but an active partner. The basis of the underlying concept of interaction mostly is that interaction occurs whenever the action of one actor is related to the action of another one. This concept is not complex enough. Thus, another important characteristic of each interaction is a common *interaction basis*. Just as humans so technical systems and humans only can coordinate their actions by referring to a common ground of knowledge. Many problems of human autonomy regarding human-technology ensembles arise from the functioning of this common interaction basis. Therefore, I want to examine it in my presentation. In doing so I rely on Hannah Arendt’s theory of action. The categories she developed especially for the realm of human interaction enable a critical perspective on the conditions of successful and failed negotiation processes within human-computer interaction.

ORGANIZING COMMITTEE

Tomislav Bracanović (University of Zagreb), *president*
Tomislav Janović (University of Zagreb)
Bojan Jerbić (University of Zagreb)
Tvrtko Jolić (Institute of Philosophy, Zagreb)
Stipe Kutleša (Institute of Philosophy, Zagreb)
Mihovil Lukić (Society for the Advancement of Philosophy, Zagreb), *secretary*
Davor Pećnjak (Institute of Philosophy, Zagreb)

ZAGREB APPLIED ETHICS CONFERENCE 2017:
THE ETHICS OF ROBOTICS AND ARTIFICIAL INTELLIGENCE

is supported by

Matica hrvatska – Matrix Croatica