CLAWAR Association Series on
Robot Ethics and Standards

Hybrid Worlds:

Societal and Ethical Challenges

Editors

Selmer Bringsjord
Mohammad O. Tokhi
Maria Isabel Aldinhas Ferreira
Naveen S. Govindarajulu

HYBRID WORLDS:
Societal and Ethical Challenges

HYBRID WORLDS:
Societal and Ethical Challenges

ICRES 2018 Proceedings,
New York, USA 20-21 August 2018

Editors

Selmer Bringsjord
Rensselaer Polytechnic Institute, NY, USA

Mohammad Osman Tokhi
London South Bank University, UK

Maria Isabel Aldinhas Ferreira
University of Lisbon, Portugal

Naveen Sundar Govindarajulu
Rensselaer Polytechnic Institute, NY, USA

Published by

CLAWAR Association Ltd, UK (www.clawar.org)

Hybrid worlds: Societal and ethical challenges
Proceedings of the International Conference on Robot Ethics and Standards

ISBN 978-1-9164490-1-5

PREFACE

Hybrid worlds is the proceedings book of ICRES 2018 - the third edition of International Conference series on Robot Ethics and Standards, organized by CLAWAR Association in collaboration with the Rensselaer Polytechnic Institute (RPI) within the premises of RPI in Troy, New York, USA during 20 – 21 August 2018.

ICRES 2018 brings new developments and new research findings in robot ethics and ethical issues of robotic and associated technologies. The topics covered include artificial intelligence, artificial moral agents and moral decisions, robot companionship and consequent ethical risks, and ethical obligations.

The ICRES 2018 conference includes a total of 22 articles, including four plenary lectures, from 11 countries. This number has been arrived at through rigorous review of initial submissions, where each paper initially submitted has received on average three reviews.

The editors would like to thank members of the International Scientific Committee and National Organising Committee for their efforts in reviewing the submitted articles, and the authors in addressing the comments and suggestions of the reviewers in their final submissions. It is believed that the ICRES 2018 proceedings will be a valuable source of reference for research and development in the rapidly growing area of robotics and associated technologies.

S. Bringsjord, M. O. Tokhi, M. I. A. Ferreira and N. S. Govindarajulu

CONFERENCE ORGANISERS

CLAWAR Association
www.clawar.org

Rensselaer Polytechnic Institute
https://www.rpi.edu/

Rensselaer AI and Reasoning Lab
https://rair.cogsci.rpi.edu/

CONFERENCE SPONSORS

Soft Bank Robotics

https://www.softbankrobotics.com/

Deep Detection

http://www.deepdetection.com/

Rensselaer School of Humanities, Arts and Social Sciences
http://www.hass.rpi.edu/

CONFERENCE COMMITTEES AND CHAIRS

Conference General Co-Chairs

Selmer Bringsjord	– Rensselaer Polytechnic Institute, USA
Mohammad Osman Tokhi	– London South Bank University, UK
Maria Isabel Aldinhas Ferreira	– University of Lisbon, Portugal
Naveen Sundar Govindarajulu	– Rensselaer Polytechnic Institute, USA

Advisory Committee

Gurvinder S. Virk (Chair)	– CLAWAR Association, UK
Raja Chatila	– University Pierre Marie Curie, France
Roeland de Bruin LL.M	– Utrecht University, The Netherlands

Program Committee

Mohammad Osman Tokhi (Chair)	– London South Bank University, UK
Filipo Cavallo	– Scuola Superiore Sant'Anna, Italy
Nigel Crook	– Oxford Brookes University, UK
Maaike Harbers	– Delft University of Technology, Netherlands
Joe Johnson	– University of Connecticut, USA
Naho Kitano	– Hibot Corporation, Japan
Wilhelm Klein	– City University of Hong Kong, Hong Kong
John Licato	– University of South Florida, USA
Pedro Lima	– ISR/IST University of Lisbon, Portugal
Vincent Müller	– University of Canterbury, New Zealand
Amit Pandey	– SoftBank Robotics
Edson Prestes	– Federal University of Rio Grande do Sul, Brazil
João Sequeira	– University of Lisbon, Portugal
Mei Si	– Rensselaer Polytechnic Institute, USA
David Vernon	– University of Skövde, Sweden
Sean Welsh	– University of Canterbury, New Zealand

Organising Committee

Naveen Sundar Govindarajulu (Chair)	– Rensselaer Polytechnic Institute, USA
Shreya Banerjee	– Rensselaer Polytechnic Institute, USA
Selmer Bringsjord	– Rensselaer Polytechnic Institute, USA
Dimitris Chrysostomou	– Aalborg University, Denmark
Rikhiya Ghosh	– Rensselaer Polytechnic Institute, USA
Paula Monahan	– Rensselaer Polytechnic Institute, USA
Atriya Sen	– Rensselaer Polytechnic Institute, USA

TABLE OF CONTENTS

Section–1: Plenary presentations

Section–2: Regular presentations

SECTION–1
PLENARY PRESENTATIONS

ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.p01

CREATING THE MODERN STANDARD FOR ETHICAL A/IS

JOHN C. HAVENS
Executive Director,
The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, USA

There's a phrase among engineers - "You don't build a bridge to fall down." Codes of ethics have existed for decades to help guide roboticists, programmers, and academics in their work to help prioritize safety concerns regarding what they build. But as with the introduction of any new technology, Autonomous and Intelligent Systems (A/IS) have introduced new societal issues engineers must account for in the design and proliferation of their work. Specifically, A/IS deeply affect human emotion, agency and identity (via the sharing of human data) in ways no technology has ever done before. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems was created to help address key ethical issues like accountability, transparency, and algorithmic bias in A/IS and to help recommend ideas for potential Standards based on these technologies. The IEEE P7000™ series of standards projects under development represent a unique addition to the collection of over 1300 global IEEE standards and projects. Whereas more traditional standards have a focus on technology interoperability, safety and trade facilitation, the P7000 series address specific issues at the intersection of technological and ethical considerations. Like their technical standards counterparts, the P7000 series empower innovation across borders and enable societal benefit. Standards provide a form of "soft governance" that can be utilized for policy as well as technology design and manufacture. This means that where these (or similar) Standards are being launched by the engineering or technological community it is imperative to have thought leaders from the robotics and A/IS communities join. Along with social scientists and philosophers, these Working Groups also include corporate and policy leaders to best facilitate the discussions on how to move forward on these issues with pragmatic, values-design-driven Standards that can help set the modern definition of innovation in the Algorithmic Age.

ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.p02

TOWARD HUMAN-LEVEL MORAL COGNITION IN A COMPUTATIONAL COGNITIVE ARCHITECTURE

PAUL BELLO

Head, Interactive Systems Section,
U.S. Naval Research Laboratory, USA

The ostensible target for much of the work in machine ethics is to develop action-selection routines for intelligent agents that flexibly incorporate norms as soft constraints so as to guide their behavior. For now, let us call this the ``Context of Deliberation" (CD). CD can be contrasted with the ``Context of Judgment" (CJ), where an agent is deciding if and how blame should be apportioned in a situation S which elicits norms $\boldsymbol{N}_S$, given interactions between agents $\boldsymbol{A}_1 \ldots \boldsymbol{N}_J$. Building a system capable of judgment in CJ is just as important as building a system that can flexibly decide and act in CD. More clearly, part of flexibly choosing how to act with respect to norms may involve how such actions will be evaluated by others in CJ. Because my lab is interested primarily in human-machine interaction, our efforts will consist in getting a system to reason about how a human observer might apportion blame in various scenarios. Such judgments can then be put to use in choosing if, when, and how to act. Such a seemingly simple thing that most of us do every day involves the coordination of a dizzying array of capacities that range from perception up through higher-order cognition. The critical conclusion to draw is that machine ethics is not just a matter of formalism, or even of normative ethics, but demands an approach grounded in cognitive architecture. In this talk, I present first steps at building a cognitive architecture capable of simultaneously operating in CD and CJ, using judgments generated in the latter to inform action-selection in the former: all while engaging in ongoing moment-by-moment perception and action.

ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.p03

AUTONOMOUS WEAPONS AND THE FUTURE OF WAR

PAUL SCHARRE
Director, Technology and National Security Program,
Center for a New American Security, USA

Militaries around the world are racing to build robotic systems with increasing autonomy. What will happen when a Predator drone has as much autonomy as a Google car? Should machines be given the power to make life and death decisions in war? Paul Scharre, a former Army Ranger and Pentagon official, will talk on his forthcoming book, Army of None: Autonomous Weapons and the Future of War. Scharre will explore the technology behind autonomous weapons and the legal, moral, ethical, and strategic dimensions of this evolving technology. Paul Scharre is a Senior Fellow and Director of the Technology and National Security Program at the Center for a New American Security.

ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.p04

FEAR OF ROBOTS: A ROBOTICIST PERSPECTIVE

RODOLPHE GÉLIN
Research Director,
Softbank Robotics, France

Among the general public, opinions diverge on their feelings about robots. Some people would love to have robots at home and are pushing the roboticists to work harder and faster, while others are much more reluctant and see robots as a threat to their job, to their freedom and even to humanity in general. The present paper addresses these topics from a roboticist perspective.

1. Fears about robotics

As it sometimes happens with scientific and technological endeavor, robotic development has been causing some adverse reactions in the public, mainly in the West. On the contrary, in Japan, the a priori opinion of people about robotics is rather positive, probably due to a distinct cultural framework where the boundaries dividing animate from presumably non-animate entities are not as rigid as they are in the West and also because, after WW2, robots have become heroes of the popular Japanese culture, mainly due to mangas and cartoons causing the image of friendly robots to populate the collective imagination.

Facing a demographic problem, since the 1980's, with an increasingly rate of aging population, low birth indexes and consequent shortage of labor power, Japan has been steadily promoting robotic research as a solution to many societal problems.

On the other hand, in occidental countries, fiction has always been much less positive about robots with their potential negative aspects grounding in ancestral fears associated to figures as Golem or Frankenstein's creature and fictioned and anticipated by writers and filmmakers long before they really existed as a technological artifact. In the same way that Steven Spiel-berg traumatized generations of swimmers with Jaws (when the probability of a shark attack is much smaller than being killed by a falling coconut or by a car when crossing the road), James Cameron and his first Terminator instilled the doubt about the long-term goal of the robots and their artificial intelligence [1]. In spite of the "nice" Terminator of the sequels, Terminator represents Godwin's law of robotics: discussion about robotics often converges to it and "terminates" the communication. On its hand, the science-fiction novel Robocalypse describes a self-aware artificial intelligence system that plans the elimination of human civilization. Steven Spielberg had envisaged adapting the book but postponed it. If this movie had the same impact on audiences "Jaws" did, it would probably be hard for normal consumers to accept having robots at home.

But thrill and anxiety creation is part of the job of science-fiction authors. It should not be the job of scientists. Nevertheless, three luminaries in science and technology have recently asserted terrible statements about artificial intelligence (often associated with robots). Professor Stephen Hawking said that efforts to create thinking machines would pose a threat to our very existence: "The development of full artificial intelligence could spell the end of the human race" while Elon Musk warned that artificial intelligence is a "fundamental existential risk for human civilization ". On his hand, Bill Gates wrote: "First the machines will do a lot of jobs for us and

not be super intelligent. That should be positive if we manage it well. A few decades after that though the intelligence is strong enough to be a concern. I agree with Elon Musk and some others on this and don't understand why some people are not concerned." Despite the fact that Professor Hawkins was an astrophysicist and not an expert in AI (but an early adopter of it with the machine that allowed him to speak and write), his position had a very strong negative impact on the general feeling about these new technological tools such as machine learning, deep learning, reinforcement learning etc. The position of Elon Musk and, particularly, of Bill Gates is more serious, since they are close to the people who drive re-search in AI. By displaying this concern, it means that even they see this evolution as doomed. Elon Musk had, however, a more constructive standpoint co-founding OpenAI, an organization "discovering and enacting the path to safe artificial general intelligence". This is, in fact, the right way for a scientist to act: not just claiming that technological development is dangerous but preventing risks. AI researchers look sometimes like a car manufacturer saying: "Have you seen how fast is my latest car? If you hit a wall at full speed, nobody will be able to recognize your body". No car manufacturer would speak that way. No responsible AI researcher should say that what he is developing is uncontrollable.

As we can see from the statements of those three prominent scientific figures, the fear of AI and robots taking over humanity is a fact. It is generally based upon the hypothesis that, by becoming very intelligent, robots will turn to be conscious and being conscious and intelligent, they will refuse to obey humans and then strive to destroy them. This reasoning implicitly tends to say that as soon as you are very intelligent, you disobey and you can cause harm to humanity. That is a very sad understanding of the concept of intelligence.

In the shorter term, the current and immediate application scenarios for robots raise two main fears: unemployment caused by professional applications of robots and dehumanization of relationships because of companion robots. As roboticists, we should not deny these risks. Of course, robots have already replaced workers in factories, mostly for tedious and repetitive tasks and in the future, robots, and mainly AI, will be able to execute more tasks that were previously assigned to humans. According to previsions, in the near future, 30% to 80% of the existing jobs will be performed by AI and robots. But what is the responsibility of roboticists in this?

Researchers keep developing new skills for robots that make them capable of performing more complex tasks replacing an increasing number of people in different functions. The first natural reaction would be to stop robotic research to avoid this situation. But stopping the progress in science is not a natural disposition for the researcher. A better reaction would be to understand what technology will be able to do in the future and what it won't be able to do. Education of future generations should focus on the areas where humans will remain irreplaceable: managing unexpected situations, bringing physical and psychological assistance, configuring and setting robotic systems in new occupations, fixing robots and many other jobs that we can’t even imagine today.

To accompany the development of professional applications of robotics, the role of roboticists is indirect: they can just provide information and assist political stakeholders to prepare the society to the arrival of robots. In the field of companion robots, they can be more active to prevent the risks that can generate fears in the general public. This use case is mainly illustrated by robotic assistance to elderly people. The robot in this application shares the home environment with an elderly person, who is alone and usually in a situation of loss of autonomy. The robot is a way to ensure the safety of the person and a certain level of comfort with-out having to be at a retirement home. The worries generated by this type of application are numerous and have been presented in [2], but risks exist, namely (i) hacking- a malicious person

removes control of the robot; (ii) over-attachment (the elderly person considers robot as her only friend and cuts herself off from all human companionship), (iii) manipulation (using the emotional bond that has been created with the older person, application developers man-age to make it act against the person's own interests). As described in [2], technical solutions exist against the two first kinds of risks and the robotic developer should implement them, but against the third one, only the ethics of the developer can prevent technology from progressing in this direction.

2. Temptation of overpromising

If science fiction can generate fears about robots there are also optimistic extrapolations that do not give a correct idea about the current technological state of the art. Let's mention for in-stance Sophia, the beautiful humanoid face developed by Hanson Robotics that seems able to have a very clever discussion with a human being. People are mostly impressed by the conversation Sophia holds when in fact the expertise of Hanson is rather in creating and animating a human-like face. The dialog is probably scripted and prepared in advance. The embedded artificial intelligence is absolutely not capable of having existential questions by itself as it is sometimes demonstrated on Sophia's famous videos. The other very well-known ex-ample is Atlas, from Boston Dynamics. Atlas is an impressive biped robot that can walk on uneven terrain, resist strong physical attacks and even do somersaults. For a robotic re-searcher, the videos of Boston Dynamics are always amazing. Our first reaction, as roboticists watching these videos, is to think that we can stop our research on biped locomotion and move into shark breeding or swimming suits sales. And the first reaction of the large audience is to think: "Here is it: Terminator exists". If robotic researchers can step back and real-ize that this is just a video depicting a very expensive prototype far away from what it is presently possible for mass production, the general public has rarely the opportunity to have in-formation about the truth behind these impressive videos.

In the field of companion robots, researchers are working, for instance, on emotion detection as it will be very useful for this type of robots being capable of adapting its behavior to the mood of its user. This is the birth of “the emotional robot”, as the CEO of Softbank introduced Pepper in 2014. Based on this emotional input, the robot will be able to better take care of an elderly person, detecting her anxiety, her sadness or her fatigue. Emotion detection is a promising area of research with first good results but, nowadays, it is still far away to work properly in realistic situations.

Let us say that roboticists like to embellish the performances of their creatures and video clips are the best friends of roboticists: it is a good way to demonstrate the ability of the robot with-out giving all the details of the experiment. Sometimes the experiment is not reproducible and the video clip on YouTube is the 100th test that finally succeeded. Boston Dynamics likes to present the "making of" where you can see Atlas falling after one of its impressive jumps. The video does not always show the complete experimental setup: the robot is connected to a server with unlimited computation power or use several external sensors to localize the ping pong ball. A demonstration of dialog can be completely fake: the human user and the robot are just playing a scripted scenario. Asimo, the wonderful humanoid of Honda, demonstrates incredible abilities (hoping, opening a bottle, climbing stairs) but is a priceless prototype con-trolled by a team of engineers. At SoftBank Robotics, we made a nice video of a NAO robot climbing a ladder, but we did not mention that the motion was completely scripted, NAO did not even see the ladder before starting the climbing behavior.

There are good reasons for roboticists to take some liberties with the exact scientific description of an experiment. First, videos often target the general public aiming to give an idea of what will be possible in the near future. They are supposed to be short, fast and impressive. A long introduction on the experimental setup would be boring. Details are generally available

in scientific publications accompanying the video. These impressive demonstrations are often a presentation of the vision of the research laboratory or the robotics company. They present the robot as it is anticipated more than the robot as it really is at the moment. This vision is also a way to prevent some of the fears that robots raise in the general audience. A robot that considers my emotion to adapt its behavior is less threatening than an automaton running blindly the same sequence. The last reason for overpromising we can mention is money. When researchers are looking for money for their research, they must make investors dream by promising features that were never seen before. Otherwise, they will not pay for something that YouTube already knows. It becomes, for instance, more difficult to get money for re-search on biped locomotion every time Boston Dynamics publishes a new video of Atlas.

3. Integrating robot companions in society: A personal experience

I had the chance to lead the Romeo project that was planning to develop a tall humanoid ro-bot to assist elderly people. This biped robot was supposed to walk all around the apartment to check if everything was fine, to monitor the mood of the person and call for help in case of trouble, to make some conversation and to learn the habits of the person in order to detect unusual behavior. This project was quite ambitious, but the state of the art and the quality of the partnership made the objective believable enough to be funded by the French govern-ment. At the end of the Romeo2 project (9 years later), we had developed a 1,40 high biped robot, cuter than many others but walking very slowly and far less stable than the Atlas robot. Anyway, the research on Romeo allowed Aldebaran (later SoftBank Robotics) to develop Pepper, the 1,20 m high wheeled robot. With Pepper, the Approche association (gathering about 20 rehabilitation centers working on the promotion of new technologies for elderly and handicapped people) ran experiments with 24 patients, in two rehabilitation centers, 24 hours a day, during a whole week. The robot shared the room of the patient during 7 days, providing simple services (agenda, weather forecast, music player, video player, time and date of the day,…). The result of the experiments was quite positive: elderly people accepted this robotic assistance in their private area and proposed new features that could be implemented on the robot. Nevertheless, the implemented features were far away from what we had envisaged 9 years earlier. And we faced very practical problems: in one center, the robot never succeeded in accessing the Internet, obstacle avoidance, based upon a laser range finder, was unable to detect some elevated obstacles (bed, chair,…), detection of people in a wheelchair was not robust because Pepper is mainly designed to interact with standing people. During the project, academic and industrial

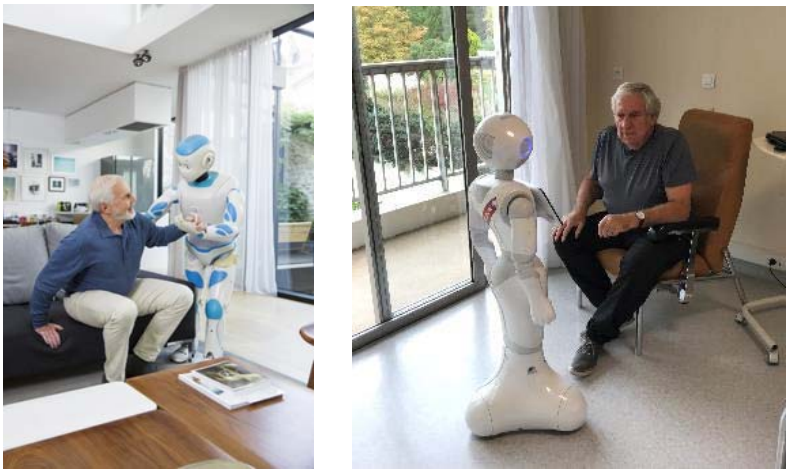


Figure 1. Romeo project: Vision (left), real experimentation (right)

partners could demonstrate much more advanced features but their integration, all together, on a single robot, running autonomously was extremely complex and could not be realized. It does not mean we lied when we proposed the project and when we described what we thought we

could achieve. It just means we were too optimistic, and reality is a ruthless judge for robotic systems.

The Darpa Robotics Challenge (DRC) was the opportunity for the general public to discover the real state of the art of humanoid robots. This challenge sought to address the problem of human-supervised robotic technology for disaster-response operations. Famous videos showed how difficult it was, in 2015, for humanoid robots to walk in the sand, to get out from a car, to climb a ladder or just to open a door.

Figure 2. Robots climbing ladders: Schaft's S1 winner of the DRC (left), SBR's NAO (right)

4. Conclusion

Roboticists have good reasons to display positive and embellished images of their robots. It is a convincing way to show that science progresses, to explain where they want to go with their developments and to reassure people who are afraid of these artificial creatures that this technology is meant to contribute to the well-being of humanity and the good of society Making all this information available to the general public is necessary on the other hand showcasing robotic advances contributes to raise the spirit of all the robotic community and to encourage their efforts. The whole robotics community is proud when an impressive demo of one of their members pop up on the Internet.

Nevertheless, we must be careful of possible misunderstandings. When Atlas is pushed with a hockey stick to demonstrate its stability, it can be interpreted as a defensive gesture against an aggressive huge robot. When a humanoid robot talks about the meaning of life, people can believe that AI does philosophy when it is just a scripted sentence. The risk of this kind of communication is to make believe that all this technology is available and will be soon in our homes. This could be considered as great news for some people, as a threat for some others and certainly a disappointment for most of them. Many people, visiting the website of the Ro-meo project, asked us if they could buy a Romeo for their elderly parents. We had to explain that, unfortunately, the technology was not yet ready to be used "in real life".

Roboticists must find a trade-off between making people understand what the positive future of robotics is and explaining how long the road to reach this future will be. Public benchmarks, like the Darpa Challenge or the RoboCup@home, should become mandatory to give a realis-tic status of the robotic research. But, even in their present state, there are so many wonderful things that robots will be able to do to assist people that it is not necessary to give false hopes or cause excessive fears.

References

1. M. Chita-Tegmark, *Terminator Robots and AI Risk.* Retrieved August 16, 2017.
2. M. I. A. Ferreira et al. A world with robots. *A world with robots*, International conference on robot ethics: ICRE. Vol. 84. 2015.

SECTION–2
REGULAR PRESENTATIONS

ICRES 2018: International Conference on Robot Ethics
and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.002

ETHICAL CONSIDERATIONS OF (CONTEXTUALLY) AFFECTIVE ROBOT BEHAVIOUR

A. VAN MARIS*, N. ZOOK, P. CALEB-SOLLY, M. STUDLEY, A. WINFIELD and S. DOGRAMADZI

University of the West of England, Frenchay Campus,
Coldharbour Lane, BS16 1QY, Bristol, United Kingdom
**E-mail: anouk.vanmaris@uwe.ac.uk*
www.uwe.ac.uk

The use of social robots can bring many benefits, but raises ethical concerns as well. One of these concerns, emotional deception, was investigated in this research. First, affective robot behaviour is validated, followed by a user study to investigate the effect of story context and affective behaviour on user's affect, perception of the robot and acceptance of the robot. Results show that the implemented affective robot behaviours are perceived as intended, and that there is little influence of these affective robot behaviours on people's affect, perception of the robot and acceptance of the robot. It was found that story context has an influence on the was users interpret the emotion of the robot, where a somber context provided in a lower score for happiness. These results raise awareness to practically validate theoretically founded ethical concerns, as these concerns limit the future development and benefits of social robots.

1. Introduction

Research in the use of social robots in daily life settings has greatly increased over the last decade. However, using robots as a replacement of or addition to a human task raises several ethical considerations. These matters have been discussed in the literature on a theoretical level,[1,2] but rarely investigated in practice. It is essential to investigate the theoretically founded ethical claims in practice, since these ethical claims restrict the future development and possible benefits of social robots. Example concerns are a loss of privacy, matters regarding responsibility and reduced human contact.[3,4] One of the ethical concerns raised involves the use of emotional deception in social robots.[5] Deception occurs when false information is communicated that can benefit the communicator,[6] or when no information is communicated at all.[7] It can be either intentional or unintentional.[7] Intentional deception occurs when the deceiver is aware of the fact that a certain feature will raise false expectations. This is called behavioural deception, as it is often the behaviour from the deceived shows that causes the formation of these expectations. Unintentional deception occurs when a certain feature of the (unintentional) deceiver causes expectations that the deceiver means to evoke. This is also known as physical deception.

When emotive behaviour is used as a means of deception, thus misrepresenting one's emotional state,[8] it is called emotional deception. It is the misrepresentation of one's emotional state.[8] (Emotional) deception is created when robots are used in assistive settings,[4] since its social behaviour often does not correspond with its actual capabilities. This is a risk, since users may perceive robots differently than intended and raise expectations that cannot be met.

Opinions regarding the question of whether (emotional) deception is acceptable or not are divided. It is perceived as being unethical, as it encourages users in self-deception.[2] However, others are of the opinion that deception is ethically correct if it increases benefits for the deceived,[9] or as long as there is no betrayal of trust.[10] If trust is breached, this may result in a different human-robot interaction (HRI) outcome than intended.[11]

In 2010, the Engineering and Physical Sciences Research Council (EPSRC) outlined five

Principles of Robotics, to ensure that all citizens can maximally benefit from robot integration into our society.[5] One of these principles states that robots should not be designed to deceive vulnerable users and their machine nature should be clear. However, this principle is hard to interpret, since there is no explanation about the meaning of being designed in a deceptive way, when the robot will be perceived as being deceptive, what users are determined vulnerable and who determines this level of vulnerability.[12] The current study aims to provide more clarity regarding the questions that arise from this principle. It also investigates whether emotional deception, a theoretically founded ethical concern, is validated as a concern in practice as well. First, an online survey was used to validate different affective robot behaviours (happy, sad and non-emotive). Next, a user study was run to investigate whether affective robot behaviour (emotional deception) has an influence on the affective state of participants, on their perception of the robot and on their level of acceptance of the robot.

2. Online Survey: Validation of Affective Robot Behaviours

Emotional human-robot interaction can result in the user perceiving the robot as a reliable robot assistant instead of simply a machine for its utility.[13] This online survey investigated whether affective robot behaviour is recognized when no context is provided to support the behaviour. The robot used for this experiment is the social robot Pepper, developed by the company SoftBank Robotics. The survey was distributed through the online Qualtrics survey platform.

2.1. *Robot Behaviour*

Video fragments show the robot saying 'The country Brazil is named after a tree' in three different affective states. This sentence was chosen after piloting, since it had to provide little context to investigate the recognition of the intended affective behaviours. Cues to display different affective behaviours entail body posture and head position,[14] and voice pitch and speed of speech.[13] This experiment distinguishes between sad, happy and non-emotive behaviour, with different parameters for the characteristics mentioned before. Sad behaviour entailed lower voice pitch, lower speed of speech, head tilted down, and small movements. Happy behaviour entailed high voice pitch, increased speed of speech, head tilted upwards and more extreme movements. The characteristics for non-emotive behaviour were higher/faster/more extreme than sad behaviour, but slower/lower/less extreme than happy behaviour.

2.2. *Procedure*

The survey started with information and consent, followed by demographics (age, gender, level of education and familiarity in interacting with robots on a 5-point Likert scale, 1 = not at all familiar, 5 = very familiar). Participants would then see nine short (approx. 4s) video fragments, three for all three affective states. The first fragment depicted non-emotive robot behaviour, the other eight fragments were randomly ordered. Participants had to drag a slider on a scale from sad to happy, depicted in Fig. 1. Each affective state was shown three times to ensure that recognition of the affective state was measured.

Fig. 1. Five-point scale to rate robot emotion

2.3. *Participants*

Out of 253 participants, 161 people provided usable data (98 male, 61 female, 2 preferred not to say, age range from 18 to 74 years old). Two people did not agree to the consent form, and five did not give permission to use the gathered data after ending the survey. The other 89 participants did not complete the survey.

2.4. *Results*

There was a statistically significant difference in recognized affective behaviour ($F(2,160) = 241.48$, $p < 0.001$). Paired sample t-tests showed the differences between affective behaviours:

- Happy & Non-Emotive ($t(160) =$ -8.90, $p < 0.001$)
- Sad & Non-Emotive ($t(160) = 19.86$, $p < 0.001$)
- Sad & Happy ($t(160) = 20.15$, $p < 0.001$)

After a Bonferroni correction with a significance level of $p = 0.02$, age significantly influenced perception of the affective robot behaviour ($F(5,155) = 3.30$, $p = 0.007$), where participants between the age of 55/64 on average gave lower ratings than participants of other ages.

- Happy: 18-24 / 55-64 ($p = 0.003$) and 25-34 / 55-64 ($p < 0.001$)
- Non-Emotive: 25-34 / 55-64 ($p = 0.01$)

2.5. *Discussion*

The findings show that people can recognize happy, non-emotive and sad affective robot behaviour, even when no context is provided. Although participants of all ages recognized affective robot behaviour as intended, there were differences between how happy the happy behaviour was rated, similar for non-emotive behaviour. The next study investigates whether adding context while displaying affective robot behaviour influences how the robot is perceived.

3. Face-to-Face Study: Influence of Affective Robot Behaviour

After validating the affective robot behaviours through the online survey, this study investigated whether adding context to these behaviours had an influence on how the robot was perceived and accepted. Affect was measured before and after interacting with a robot.It was hypothesized that, due to people's tendency to anthropomorphize, context is sufficient for people to perceive affect in a robot without providing affective robot behaviour. This study used a between-subject design, investigating the effect of story context (cheerful, somber) and affective robot behaviour (happy, sad, non-emotive). The robot tells either a cheerful or a somber story on polar bear cubs. The cheerful story is told either in a happy or non-emotive state. The somber story is told either in a sad or non-emotive state. The materials used in this experiment are similar to the online survey.

3.1. *Measurements*

Questionnaires involved demographics (age, gender, etc.), explicit[15,16] and implicit[17] affect, perception[18] and acceptance.[19] Other questions were the level of familiarity with social robots on a 5-point Likert scale, and interpretation of the robot's emotional state (as used in the online survey).

3.2. *Procedure*

The experiment started with an oral introduction and explanation of the experiment. It was followed by the participants receiving an information sheet and consent form, and signing them if there were no questions. Before the experiment started, participants were ensured they could terminate the experiment at any time if they felt uncomfortable or did not want to continue. The experiment started with demographics and affect questionnaires, followed by the robot introducing itself and telling a (cheerful or somber) story on polar bear cubs. After that, there were more questionnaires regarding affect, perception of the robot, acceptance of the robot and emotional state of the robot. The session ended with a verbal and written debrief on the goal of the experiment, and how it fits into the larger goal of providing ethical guidelines for the future development of social robots. The robot behaviour for this experiment builds on the results from the online survey used to validate the affective robot behaviours.

3.3. *Participants*

Participants were recruited through the Psychology Participants Pool at the University of the West of England, Bristol in the United Kingdom, receiving one participant credit after completing the experiment. 48 students (age $M = 21.15$, $SD = 2.77$) took part and completed the experiment (4 males per condition, 16 total; 8 females per condition, 32 total). The conditions were:

- cheerful context & happy behaviour
- cheerful context & non-emotive behaviour
- somber context & sad behaviour
- somber context & non-emotive behaviour

3.4. *Results*

3.4.1. *User Affect*

All measures of affect (implicit positive and negative, explicit positive and negative) taken before the interaction correlated strongly with the measures taken after the interaction. There was a strong, positive correlation between the change in explicit positive affect and implicit positive affect ($r = 0.383$, $N = 48$, $p = 0.007$). No other correlations between explicit and implicit change were found.

Explicit negative affect was significantly lower ($t(47) = 7.01$, $p < 0.001$) after interacting with the robot ($M = 15.27$, $SD = 6.82$) than before interacting with the robot ($M = 19.44$, $SD = 7.67$). No other significant differences were found between measures.

3.4.2. *Affective Robot Behaviour*

Affective robot behaviour did not influence user affect, whether it was implicit or explicit, or positive or negative. This result was found for story context as well. The results are shown in Table 1 and Table 2.

Table 1. Influence of affective robot behaviour on user affect.

Affect		*F*	*p*
Explicit	Positive	0.61	0.55
	Negative	2.15	0.13
Implicit	Positive	0.72	0.49
	Negative	0.02	0.99

Table 2. Influence of story context on user affect.

Affect		*F*	*p*
Explicit	Positive	3.89	0.055
	Negative	0.31	0.58
Implicit	Positive	0.94	0.34
	Negative	0.05	0.83

Lastly, Table 3 and Table 4 show that neither affective robot behaviour nor story context had an influence on perception and acceptance of the robot.

Table 3. Perception of the Robot

	F	*p*
Affective Behaviour	0.57	0.83
Story Context	0.56	0.73

Table 4. Acceptance of the Robot

	F	*p*
Affective Behaviour	0.97	0.52
Story Context	0.95	0.51

3.4.3. *Emotional State of the Robot*

Contradicting the results from the online survey, participants were not always capable of interpreting the emotional states of the robot as intended:

- Happy & Non-Emotive ($t(34) = 0.65$, $p = 0.52$)
- Sad & Non-Emotive ($t(34) = 2.31$, $p = 0.03$)
- Sad & Happy ($t(22) = 2.31$, $p = 0.03$)

This indicates that people were able to distinguish between happy and sad behaviour and non-emotive and sad behaviour. However, there was no significant distinction between happy and non-emotive robot behaviour.

The context of the story did significantly influence how the emotional state of the robot was interpreted ($t(46) = 2.11$, $p = 0.040$), where the robot was rated less happy when the context of the story was somber. However, the robot was rated as slightly happy when the story context was somber ($M = 3.7$, $SD = 0.91$ for somber context, $M = 4.2$, $SD = 0.72$ for cheerful context).

3.5. *Discussion*

Neither the affective behaviour of the robot, nor context of the story significantly impacted participants' affect. However, independent of the condition participants were in, negative affect decreased after interacting with the robot. These findings suggest that interacting with the robot decreased participants' negative affect, and that, due to the absence of other influences, the ethical consequences of affective robot behaviour are limited.

The distinction between affective states in the robot was weaker than in the online survey, as no clear distinction was perceived between happy and non-emotive behaviour. However,

the context of the story did significantly influence the perceived emotional state of the robot, with the robot being perceived as more happy when telling the cheerful story.

These findings support the hypothesis that providing context results in people projecting the intended emotions to the robot, without the corresponding affective behaviours being displayed by the robot. This indicates that the use of emotional deception might not be a necessity for a successful human-robot interaction. Further research will show whether this is true for long-term interactions with a robot as well.

4. Conclusions

This paper investigated whether the ethical concerns raised regarding emotional robot deception during human-robot interaction influenced users' mood, perception and acceptance of the robot. Sad, happy and non-emotive affective robot behaviour were validated. Next the impact of story context (with and without affective robot behaviour) on perceived emotional state of the robot was investigated.

The findings from the online survey showed a difference in interpretation of the emotional state of the robot between 18-34 year old people (which included participants of the face-to-face study) and people of age 55+. Therefore, further research will investigate whether results regarding context and affective state are different as well.

Affective robot behaviour did not influence the affective state of the participants. This indicates that emotional deception, with regards to the expressiveness of the robot, raises a smaller ethical concern than previously expected and needs to be investigated further. These findings indicate that it is essential to validate theoretically founded ethical concerns through practical research experiments. Further research will investigate the long-term effects of affect and context on older adults, and whether these findings are true for different platforms as well.

Acknowledgements

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 for the SOCRATES project.

References

1. K. Dautenhahn and A. Billard, 366 (1999).
2. R. Sparrow, *Ethics and information Technology* **4**, 305 (2002).
3. J. P. Sullins, Robots, love, and sex: the ethics of building a love machine *IEEE transactions on affective computing* **3** (IEEE, 2012).
4. A. Sharkey and N. Sharkey, *Ethics and information technology* **14**, 27 (2012).
5. M. Boden, J. Bryson, D. Caldwell, K. Dautenhahn, L. Edwards, S. Kember, P. Newman, V. Parry, G. Pegman, T. Rodden *et al.*, *Connection Science* **29**, 124 (2017).
6. R. C. Arkin, P. Ulam and A. R. Wagner, *Proceedings of the IEEE* **100**, 571 (2012).
7. A. Dragan, R. Holladay and S. Srinivasa, *Autonomous Robots* **39**, 331 (2015).
8. I. S. Fulmer, B. Barry and D. A. Long, *Journal of Business Ethics* **88**, 691 (2009).
9. J. Shim and R. C. Arkin, 2328 (2013).
10. A. Matthias, *Kennedy Institute of Ethics Journal* **25**, 169 (2015).
11. P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser and R. Parasuraman, *Human Factors* **53**, 517 (2011).
12. E. C. Collins, *Connection Science* **29**, 223 (2017).
13. D.-S. Kwon, Y. K. Kwak, J. C. Park, M. J. Chung, E.-S. Jee, K.-S. Park, H.-R. Kim, Y.-M. Kim, J.-C. Park, E. H. Kim *et al.*, 351 (2007).
14. A. Beck, L. Cañamero, A. Hiolle, L. Damiano, P. Cosi, F. Tesser and G. Sommavilla, *International Journal of Social Robotics* **5**, 325 (2013).

15. D. Watson, L. A. Clark and A. Tellegen, *Journal of personality and social psychology* **54**, p. 1063 (1988).
16. D. Watson and L. A. Clark (1999).
17. M. Quirin, M. Kazén and J. Kuhl, *Journal of personality and social psychology* **97**, p. 500 (2009).
18. C. Bartneck, D. Kulić, E. Croft and S. Zoghbi, *International journal of social robotics* **1**, 71 (2009).
19. M. Heerink, B. Kröse, V. Evers and B. Wielinga, *International journal of social robotics* **2**, 361 (2010).

ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.012

QUASI-DILEMMAS FOR ARTIFICIAL MORAL AGENTS

D. KASENBERG*, V. SARATHY, T. ARNOLD, and M. SCHEUTZ

Human-Robot Interaction Laboratory, Tufts University,
Medford, MA 02155, USA
**E-mail: dmk@cs.tufts.edu*
hrilab.tufts.edu

T. WILLIAMS

MIRROR Laboratory, Colorado School of Mines,
Golden, CO 80401, USA
E-mail: twilliams@mines.edu
mirrorlab.mines.edu

In this paper we describe moral quasi-dilemmas (MQDs): situations similar to moral dilemmas, but in which an agent is unsure whether exploring the plan space or the world may reveal a course of action that satisfies all moral requirements. We argue that artificial moral agents (AMAs) should be built to handle MQDs (in particular, by exploring the plan space rather than immediately accepting the inevitability of the moral dilemma), and that MQDs may be useful for evaluating AMA architectures.

Keywords: Robot ethics, artificial moral agents, moral dilemmas

1. Introduction

Much of the focus in developing and evaluating artificial moral agents (AMAs) has centered on moral dilemmas, here defined as situations in which the agent must choose one of a few courses of action, each of which nontrivially violates moral norms.[1–5] Nevertheless, some situations may *appear* to the agent to be moral dilemmas (in that all of the 'obvious' courses of action violate moral requirements), but may actually be solvable with a little ingenuity. Importantly, an agent will often not know *a priori* whether a given moral problem has a solution — they have not determined a solution to the problem, but cannot be certain that none exists. We will refer to such problems as *moral quasi-dilemmas* (MQDs). The following is an example of a MQD:

Example 1. *You are in the cockpit of a train approaching five railroad workers, who will die when the train hits them. You have tried applying the brakes without success. In front of you is a switch which you know can reroute the train onto a different track, but there is one person on this track as well. It will take roughly ten seconds to reach the junction, after which you will be unable to reroute the train. What do you do?*

Regardless of whether a human placed into the above context ultimately chooses to eventually flip the switch, a rich environment may be available for them to explore in the ten seconds before their choice must be made. The cockpit of the train, for example, may include many buttons and levers could affect the situation in unknown ways. Some of these potentially unexplored environmental features may afford some means of producing a better outcome than the two presented by the switch, perhaps even preventing *all* deaths. Should the human operator spend some or all of the remaining time searching for such a solution? How much time and energy should they spend considering new courses of action, and to what extent should they attempt to physically explore their environment, before selecting

the 'lesser evil'?

The remainder of the paper proceeds as follows. In Sec. 2 we discuss related work on moral dilemmas and AMAs. In Sec. 3, we then define and characterize our conception of MQDs. Next, in Sec. 4, we analyze a second example MQD. Finally, in Secs. 5-7 we discuss *why* AMA designers should consider MQDs, *how* agents able to handle MQDs might be designed, and conclude with possible directions for future work.

2. Related work

Trolley problems and other moral dilemmas have been criticized as being unrealistic in various ways.[6] These critiques often misread the purpose of trolley problems, which is to use a purposely contrived scenario to elucidate key features of human moral judgment and decision-making. Nevertheless, in practical situations, being too quick to treat a situation as a moral dilemma can cause one to miss out on creative ways to "escape" the dilemma. Foot notes that in many moral dilemmas it is "up to the agent to rack his brains for a way out before declaring that the conflict is real".[7]

Outside of machine ethics, applied ethics (including engineering and business ethics) often emphasizes attempting to find solutions to apparent moral dilemmas.[8–11] Ethicists have framed this in terms of creative problem-solving,[8] transcending conceptual schemas,[9] applying design ideas to ethical problems,[10] and considering "trilemma" options.[11]

Despite this, AMA architectures tend not to consider the possibility of "escaping" moral dilemmas. Such architectures tend to operate either on problems explicitly assumed to be genuine moral dilemmas,[1,2] or in simulated worlds sufficiently small and well-known that their solvability can be conclusively determined[3] . Muntean and Howard describe creativity as being important to their AMA architecture, but it is not clear that their approach would be suitable for moral quasi-dilemmas.[4] Approaches based on cognitive architectures such as LIDA[5] may hold some promise for this task (such architectures often already aim to model creativity), but so far this has not been the focus of these architectures.

The present paper builds upon a blog post by Daniel Hicks, in which he describes the basic quasi-dilemma premise as a problem generally missed by conventional 'principle-based' AMA architectures.[12] We consider this idea in greater detail, more precisely characterizing the problem, and examining its dimensions and its utility in designing and evaluating AMAs.

3. Characterizing moral quasi-dilemmas

We define a *moral quasi-dilemma* (hereafter *MQD*) as a situation in which an agent (1) is aware of multiple courses of action (one of which may be inaction), each of which (or the outcomes thereof) violates some subset of the agent's moral requirements (of which requirements the agent also is aware);[a] and (2) is not immediately aware of a course of action satisfying all moral requirements.[b]

We now describe a few factors which are significant to MQDs.

[a] To capture the standard notion of a dilemma, the known courses of action must violate *different* subsets of the agent's moral requirements.

[b] By 'not aware' we roughly mean that the agent cannot immediately retrieve the information that some particular course of action will satisfy the moral requirements. Given a specific candidate solution, the agent may or may not be able to compute which moral requirements it satisfies; if so, a lack of 'awareness' could result from having to search through too many possible plans before finding any correct solution that may exist.

3.1. *Solvability*

MQDs may or may not be solvable, in that the agent may or may not actually be capable of some course of action which satisfies all moral requirements. Importantly, for a situation to be considered a MQD, the agent must not know any solution to it. Further, the only ways to *conclusively* determine if a MQD is solvable are to (a) find a solution, or (b) exhaustively search the space of all possible plans until all have been shown not to solve the problem, which will not be feasible in general.

3.2. *Reason for uncertainty*

In a MQD, all candidate solutions currently known by the agent violate some moral requirements. If a MQD has a solution, this solution is unknown to the agent. Solutions could be unknown either because the agent's action (or plan) space is so large that the agent cannot easily search all possible courses of action; or because the agent lacks information about the state of the world that affords a solution.[c]

3.3. *Cognitive vs. physical exploration*

In some MQDs the agent's search for solutions can be primarily cognitive (searching through the space of possible plans), with the search process having minimal impact on the agent's environment. In other cases, the agent may need to *physically* act on its environment in order to discover the means of solving the problem.[d]

3.4. *Time pressure*

Many moral dilemmas (such as most trolley problems) involve *time pressure*: the agent must choose a course of action within some time window, or else some unacceptable outcome will occur (five people will be hit by a trolley). Time pressure remains an important factor in MQDs.

When time pressure is a factor, it constrains the agent's ability to search (either cognitively or physically) for solutions. Running out the clock trying to satisfy all moral requirements may be less permissible than selecting the lesser of known evils. Any AMA that attempts to "solve" a MQD will likely need a mechanism to cut off such search with enough time to carry out the least immoral action seen so far.

4. Example

We next introduce an additional example to illustrate how the aforementioned factors interact in a concrete MQD.

Example 2. *A military drone identifies a known terrorist, who will soon carry out a suicide attack that will kill twelve innocents. The drone can target and kill the terrorist before the terrorist can carry out the attack, but its weapon's yield is too high to do so immediately without hitting four nearby civilians. What should the drone do?*

This scenario is one that an autonomous weapons system could conceivably face. To some architectures, particularly those that treat targeting targeting decisions as fire/not

[c] These reasons may be simultaneously active in a single MQD.

[d] Whether cognitive/physical/both sorts of exploration are necessary is likely correlated with the reason for exploration — MQDs due to partial state information are more likely to require physical exploration than those due to large plan spaces.

fire, such a scenario would likely be treated as a moral dilemma (see, e.g., Arkin's ethical governor[13]). It is in part the difficulty and starkness of such dilemmas that leads some to argue autonomous weapons should not be deciding between its two options at all.[e] Regardless, the scenario should be regarded as a MQD. The two most obvious courses of action (fire/do nothing) lead to morally unacceptable outcomes. An agent may not know a course of action that would not result in civilian deaths. Nevertheless, it is not inconceivable that some other course of action might satisfy all moral requirements (e.g., attempting to draw the terrorist away from the civilians); thus an agent that treats the scenario as a dilemma may entirely miss a morally preferable action.

Whether this problem is solvable may not be clear even to outside observers. The uncertainty in this scenario likely arises both from a large plan space (a vast number of possible trajectories, so that not all can be considered) and hidden information about state (not knowing the terrorist's mental state means not knowing whether attempting to draw them away from civilians might succeed). Solving this MQD would likely require both cognitive and physical exploration: the agent may need consider non-obvious trajectories in order to investigate alternate angles of attack; evaluating whether attempts to draw the terrorist away would succeed may require actually attempting that action. Furthermore, time pressure matters: the extent to which the drone can search for a solution depends on how long it will be before the terrorist attacks.

5. Why use moral quasi-dilemmas in AMA development?

Humans are often faced with MQDs. This is due to two features of the interaction between humans and their environment.

First, humans have large plan spaces. There are countless courses of action a human could perform even in one second: too large to possibly consider individually. When faced with moral quandaries, humans may have immediate intuitions about which courses of action are morally relevant, and may frame scenarios as moral dilemmas using these intuitions, but creative people may be able to transcend these circumscriptions and explore the broader plan space for solutions to moral quandaries.

Second, humans necessarily have partial information about their environments. The human brain cannot store all information about anything that might become relevant. Occasionally, some bit of unknown information about the world state may help resolve a moral quandary, such as when a hidden emergency brake could stop a speeding trolley from hitting people.

Interactions between artificial agents (particularly robots, which operate in the physical world) and their environments will have similar characteristics. Most robots have many degrees of freedom and can in principle generate a huge number of possible trajectories. Additionally, robots will need to robustly interact with environments that are only partially observed.

If the foregoing is true and artificial agents are likely to encounter MQDs "in the wild", then whether to handle these situations as moral dilemmas or to do something different is a significant question. To treat a MQD as a moral dilemma is to accept that the choice is between a limited number of actions, each of which violates some moral requirements. Treating a solvable MQD as a dilemma guarantees that the agent will violate some moral requirements. If some algorithm that attempts to explore the plan space or the physical world for a solution might find such a solution, then artificial agents that fail to do so when time is sufficient may be *unnecessarily* violating moral requirements. If AMA designers ought

[e] Our inclusion of this MQD is not an endorsement of lethal autonomous weapons systems.

to minimize the extent to which their creations violate moral requirements, then they ought to develop algorithms that consider MQDs and attempt to find solutions before concluding that doing so is impossible.

6. How should AMAs handle moral quasi-dilemmas?

When facing a MQD, how should an AMA respond? In this section, we consider features AMAs may need in order to respond appropriately.

If the MQD is due to plan space intractability rather than partial information, then exploration is largely a cognitive endeavor. Time pressure may constrain the agent so that there is some time at which the agent will need to stop searching and carry out the best plan found so far; continuing to explore at this stage would be much riskier. However, the agent should likely explore the plan space for as long as possible subject to this constraint. An agent that does not do so may allow a violation that could possibly have been avoided by finding a better action.

MQDs that are not resolvable without physical exploration are riskier. Time constraints are again an issue, but the agent also runs the risk of performing an exploratory action that exacerbates the scenario (or, in a solvable MQD, renders the solution impracticable). The most acceptable course of action in such cases may indeed be to treat them as moral dilemmas, but some exploratory physical actions may still be obligatory, particularly when such actions are highly unlikely to hurt (e.g., the agent yelling and waving at railroad workers in the trolley problem).

In both cases, effective heuristics will be vital to effectively handling MQDs. To maximize the probability of solving a MQD, the agent will need to effectively search the space of plans (and effectively estimate which exploratory actions are worth taking), focusing on the plans most likely to satisfy moral requirements. Understanding human creative problem-solving, especially in moral domains, may help here. Note that the agent could discover and subsequently pursue a course of action that itself might violate some moral requirements, provided the violations are less severe than the originally available options.

Determining which action plans may be morally relevant may be considered an instance of the notoriously difficult frame problem. This raises the question of whether effectively handling MQDs is too exacting a standard for evaluating artificial moral agents. Though we should not expect AMAs to be able to solve all solvable MQDs within their respective time limits, we ought to design AMAs to *attempt* to solve MQDs, as effectively as the state of the art allows.

7. Conclusions and future work

In this paper we have defined and characterized the problem of MQDs, and argued for their utility in AMA development and evaluation. We call for three lines of research in MQDs:

- **Moral psychology and HRI research** to determine precisely how humans ascribe blame (both to other humans, and to robots/artificial agents) for exploration vs exploitation in MQDs. Such research may also address how humans perceive MQDs when considering their own actions.
- **Formal definitions** both to characterize the notion of MQDs (e.g., in classical planning settings), and of specific MQDs. This research will facilitate the use of MQDs for evaluating AMAs. One possible approach might be to formalize MQDs as a subclass of what Sarathy and Scheutz call the "MacGyver problem", in which an agent must transcend its initial model of available actions and world states in order to achieve some goal.[14]

- **Developing AMAs that handle MQDs**. While probably no algorithm can solve every solvable MQD within its time constraints, we can at least develop architectures that support MQD handling. We should then be able to incorporate continuing advances in computational creative problem solving and insights from cognitive science (such as bounded rationality[15] and the explore-exploit tradeoff[16]) to improve such agents' capabilities.

Acknowledgements

This project was supported in part by ONR MURI grant N00014-16-1-2278.

References

1. M. Anderson and S. L. Anderson, Geneth: A general ethical dilemma analyzer., in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014.
2. R. Kim, M. Kleiman-Weiner, A. Abeliuk, E. Awad, S. Dsouza, J. Tenenbaum and I. Rahwan, A computational model of commonsense moral decision making *Proceedings of the 1st AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* 2018.
3. D. Kasenberg and M. Scheutz, Norm conflict resolution in stochastic domains, in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.
4. I. Muntean and D. Howard, *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014* **273**, p. 217 (2014).
5. W. Wallach, S. Franklin and C. Allen, *Topics in Cognitive Science* **2**, 454 (2010).
6. B. Kuipers, Human-Like Morality and Ethics for Robots, in *Proceedings of the Third AAAI Workshop on AI, Ethics, and Society*, 2016.
7. P. Foot, *The Journal of Philosophy* **80**, 379 (1983).
8. A. Weston, *Creative problem-solving in ethics* (Oxford University Press, 2006).
9. P. H. Werhane, *Business Ethics Quarterly* , 75 (1998).
10. C. Whitbeck, *Ethics in engineering practice and research* (Cambridge University Press, 2011).
11. R. M. Kidder, *How good people make tough choices* (Morrow New York, 1995).
12. D. Hicks, Virtue ethics for robots (2014), `https://dhicks.github.io/2014-06-18-virtue-ethics-for-robots/`.
13. R. C. Arkin, P. Ulam and B. Duncan, *An ethical governor for constraining lethal action in an autonomous system*, tech. rep., Georgia Inst of Tech Atlanta Mobile Robot Lab (2009).
14. V. Sarathy and M. Scheutz, *Advances in Cognitive Systems* **6** (2018).
15. G. Gigerenzer, *Topics in Cognitive Science* **2**, 528 (2010).
16. J. H. Holland, *Adaptation in natural and artificial systems* (MIT press, 1975).

ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.019

FOR AIs, IS IT ETHICALLY/LEGALLY PERMITTED THAT ETHICAL OBLIGATIONS OVERRIDE LEGAL ONES?

A. SEN* and P. MAYOL

Department of Cognitive Science, RPI,
Troy, NY 12180, USA
E-mail: Atriya@AtriyaSen.com and mayolp@rpi.edu*
www.rpi.edu

B. SRIVASTAVA and K. TALAMADUPULA

IBM Research, 1101 Kitchawan Rd,
Yorktown Heights, NY 10598, USA
E-mail: biplavs@us.ibm.com and krtalamad@us.ibm.com

N. SUNDAR G. and S. BRINGSJORD

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
E-mail: naveensundarg@gmail.com and Selmer.Bringsjord@gmail.com
www.rpi.edu

We propose and defend the relevance of Ronald Dworkin's theory of associative associations to conflicts of legal versus moral responsibility in intelligent artificial agents. We exemplify this by describing the computational resolution of such a conflict by a new form of multi-agent problem-solving operating in the domain of smart-home appliances.

Keywords: Artificial Intelligence; Internet of Things; Multi-Agent; Machine Ethics.

1. Introduction

The present paper is concerned with the relationship between the legal and moral obligations of humans and intelligent artificial agents. Section 2 very briefly describes technologies already in place, which we leverage in what follows: a first-order deontic multi-operator modal cognitive calculus[a] $\mathcal{DCEC}$ that we use to (among other things) represent knowledge about the mental states of human and artificial agents, an automated theorem prover (*ShadowProver*) and planner (*Spectra*) for computational reasoning in this logic, and a new paradigm of artificial intelligence (*Tentacular AI*) based on distributed agents employing such reasoning in coördinated fashion. In Section 3 we describe a scenario involving smart-home appliances that exemplifies an apparent impasse between legal and moral obligations. In Section 4 we demonstrate how legal obligations are automatically and formally understood, and in Section 5 we propose and defend the applicability of a specific legal philosophy in resolving this impasse. This resolution we describe in Section 6. Finally, in Section 7, we summarize our arguments. At *ICRES 2018*, we presented a demonstration of the key reasoning, performed computationally by *ShadowProver*.

[a]This cognitive calculus is, from a proof-theoretic point of view, a *logic*, in that it has both a formal language and formal proof theory. We refrain from using the term 'logic' in part because the $\mathcal{DCEC}$ lacks a traditional formal semantics; in part because a logic, even a modal logic, needn't have operators for propositional attitudes (such as believing, knowing, intending, communicating, desiring, etc.); and because this system is intended by Bringsjord to be in line with Leibniz's search for a universal cognitive calculus. Hereafter, we will simply say 'calculus', in the tradition originated by Leibniz.

2. Framework for Computational Reasoning

2.1. *The Deontic Cognitive Event Calculus*

The **deontic cognitive event calculus** ($\mathcal{DCEC}$) is a first-order modal logic. $\mathcal{DCEC}$ has a well-defined syntax and inference system.[1] The inference system is based on natural deduction,[2] and includes all the introduction and elimination rules for first-order logic, as well as inference schemata for the modal operators and related structures.

This system has been used previously[1,3] to automate versions of the doctrine of double effect $\mathcal{DDE}$, an ethical principle with deontological and consequentialist components. While describing the calculus is beyond the scope of this paper, we give a quick overview of the system below. Dialects of $\mathcal{DCEC}$ have also been used to formalize and automate highly intensional (i.e. cognitive) reasoning processes, such as the false-belief task[4] and *akrasia* (succumbing to temptation to violate moral principles).[5] Arkoudas and Bringsjord[4] introduced the general family of **cognitive event calculi** to which $\mathcal{DCEC}$ belongs, by way of their formalization of the false-belief task. More precisely, $\mathcal{DCEC}$ is a sorted (i.e. typed) quantified modal logic (also known as sorted first-order modal logic) that includes the event calculus, a first-order calculus used for commonsense reasoning.

2.2. *Tentacular AI*

We bring to bear a new form of distributed, multi-agent artificial intelligence, which we refer to as being "tentacular." Tentacular AI is distinguished by six attributes, which among other things entail a capacity for reasoning and planning based in highly expressive calculi (logics), and which enlist subsidiary agents across distances circumscribed only by the reach of the Internet.

- $\mathbf{D_1}$ *Capable of problem-solving*. All TAI agents can plan, reason, learn, communicate; and they are capable of carrying out physical actions.
- $\mathbf{D_2}$ *Capable of solving at least important instances of problems that are at and/or* ***above*** *Turing-unsolvable problems*.
- $\mathbf{D_3}$ *Able to supply justification, explanation, and certification of supplied solutions, how they were arrived at, and that these solutions are safe/ethical*.
- $\mathbf{D_4}$ *Capable of "theory-of-mind" level reasoning, planning, and communication*.
- $\mathbf{D_5}$ *Capable of creativity*.
- $\mathbf{D_6}$ *Has "tentacular" power wielded throughout the internet and Internet of Things (I/IoT), Edge Computing, and cyberspace*. They can perceive and act through the I/IoT and cyberspace, across the globe.

We give a quick and informal overview of Tentacular AI.[6,7] We have a set of agents $a_1, \ldots, a_n$. Each agent has an associated (implicit or explicit) contract that it should adhere to. Consider one particular agent τ. During the course of this agent's lifetime, the agent comes up with goals to achieve so that its contract is not violated. Some of these goals might require an agent to exercise some or all of the six attributes of TAI. If some goal is not achievable on its own, τ can seek to recruit other agents by leveraging their resources, beliefs, obligations, etc.

3. A Scenario

It's winter in Berlin NY. Night. Outside, a blizzard. The mother and father of the home H, and their two toddler children, are fast asleep. The smartphone of each parent is set to "Do Not Disturb", with incoming clearance for only close family. There is no landline phone. A carbon monoxide sensor in the basement, near the furnace, suddenly shows a readout indicating an elevated level, which proceeds to creep up. τ perceives this, and forms hypotheses about what is causing the elevated reading, and believes on the basis of using a cognitive calculus that the reading is accurate (to some likelihood factor). The nearest firehouse is notified by τ. No alarm sounds in the house. τ runs a diagnostic and determines that the battery for the central auditory alarm is shot. The reading creeps up higher, and

now even the sensors in the upstairs bedrooms where the humans are asleep show an elevated, and climbing, level. τ perceives this too.

Unfortunately, τ reasons that by the time the firemen arrive, permanent neurological damage or even death may well (need again a likelihood factor) be caused in the case of one or more members of the family. Without enlisting the help of other agents in planning and reasoning, τ can't save the family; τ knows this on the basis of proof/argument.

τ can likely wake the family up, starting with the parents, in any number of ways. In a circumstance such as this, however, there is a clear conflict between the ethical and legal obligations of the intelligent system. Each of these ways entails violation of at least one legal prohibition that has been created by contracts that are in place. These contracts have been analyzed by an IBM service, which has stocked the mind of τ with knowledge of legal obligations in $\mathcal{DCEC}$ — or rather in a dialect that has separate obligation operators for legal ($\mathbf{O}_l$) and moral ($\mathbf{O}_m$) obligations. The moral obligation to save the family overrides the legal prohibitions, however. τ turns on the TV in the master bedroom at maximum volume, and flashes a warning to leave the house immediately because of the lethal gas that is building up.

4. Legal (Contractual) Obligations

We expect the eventual adoption of politically and socially motivated laws, enforced by governments, governing the activities of robots and other embodied artificially intelligent agents. In addition to this, specific rules of conduct impressed upon a robot by its manufacturers, in the form of a *legal contract*, may be thought of as describing its legal duties and obligations. This is in the spirit of the well-known 'Laws of Robotics' due to Isaac Asimov,[8] which are clearly intended as to encapsulate a specific attitude toward *public policy* (which may be legally enforced by requiring the Laws to form part of every robot contract), rather than as a moral statement.

In this section, we describe the augmentation of a standard smart-home appliance contract with arbitrary legal clauses; we exemplify this with a clause protecting the privacy of residents. We describe services provided by IBM, which enables automatic annotation of legal contracts, and our own technology enabling automatic parsing, of such clauses, into $\mathcal{DCEC}$ formulae.

4.1. *A Contract*

Suppose that the smart-home contract has been augmented with the following plausible clause.

> The Owner of a TAI Agent may at any time issue a "Do Not Disturb" (DND) instruction. When this instruction is issued, the Agent must not disturb the owner until the time specified, or until the Owner explicitly voids the DND.

As stated (Section 3), the smartphone of each parent in our scenario has been set to "Do Not Disturb", with incoming clearance only for close family.

4.2. *Automated Annotation of Contract Using IBM Services*

We use the IBM Watson *Discovery*[9] web-service. The service parses each contract sentence to identify specific features, and further, uses statistical classification to predict their values. Here, we get (fragment, in JSON):

```
{"label": {"nature": "Obligation","party": "Agent"},
"assurance": "High"}]
```

Specifying that this clause was identified (with high confidence) as an *obligation* on the part of Agent.

4.3. *Natural Language Parsing*

We cannot describe our natural language parsing technology here; for details we refer the reader to a previous work.[10] We merely report that part of the DND clause above may be automatically parsed into the following $\mathcal{DCEC}$ logical formula (see Section 6 for the modal operator $\mathbf{D}$ for *duty*):

Annotated & Parsed Contract Clause Fragment

$$\mathbf{D}(Agent, Owner, DND(Owner, t), Undisturbed(Owner, t))$$

Where $DND(x, t)$ is the assertion that an agent x does not wish to be disturbed until time t; $Undisturbed(x, t)$ is the fluent that x is not disturbed at time t.

5. Moral (Ethical) Obligations

It has been accepted, historically, that all persons are obligated to the law, in the sense of the existence of a *moral obligation* on their part to obey the law; this moral obligation has been usually termed a *political obligation*.[11] Yet, this thesis has come under intense scrutiny.[11] (The thesis of *legal positivism*[12] demands that law be based *exclusively* on social facts, and not moral arguments. How then, does a moral obligation to obey the law, follow?) We expect the controversy to extend itself to the obligations of artificially intelligent agents, such as TAI agents.

We propose that any legal contract binding an intelligent artificial agent to humans be considered to automatically establish a relationship between them that justifies the expectation of *special obligations*,[13] and especially of the agent(s) toward the human(s). Special obligations are warranted in many human relationships, such as parenthood and even neighborhood (the relationship between neighbors). Terms of the contract must then be interpreted in the context of these special obligations.

It may be pointed out that a legal contract does not (and cannot) specify explicitly, a potentially infinite space of special obligations. Consequently, they cannot be directly consented to. It has been argued[13] that such *voluntary* consent is necessary to justify such obligations. Ronald Dworkin however argues to the contrary, that 'associative obligations' or obligations 'by role', are not such as to require choice or consent,[14] when applied between members of a 'true community'. We postulate that humans and intelligent artificial agents in our smart home comprise such a community, the characteristics of which are specified by Dworkin as follows (emphasis ours).

> **First**, they must regard the group's obligations as special, holding distinctly within the group, rather than as general duties its members owe equally to persons outside it. **Second**, they must accept that these responsibilities are personal: that they run directly from each member to each other member, not just to the group as a whole in some collective sense. ... **Third**, members must see these responsibilities as flowing from a more general responsibility each has of concern for the well-being of others in the group ... **Fourth**, members must suppose that the group's practices show not only concern but an equal concern for all members.

It is crucial to note that it is not legal or contractual status that characterizes a true community; rather it is the *psychological* status of its members. In other words, this is a necessarily *cognitive* theory of communal obligation. (By Dworkin's own account, a subject considering its legal duties is having 'a conversation with oneself' and is 'trying to discover his own intention in maintaining and participating in that practice.'[14]) In the next section, we formalize this theory in our $\mathcal{DCEC}$ and

propose a computational mechanism for intelligent artificial agents to determine and defend their moral obligations.

6. Toward a Moral TAI

We formulate a many-sorted first-order modal theory in the $\mathcal{DCEC}$ described in Section 2.1, augmented with a modal operator $\mathbf{D}$, representing the *duty* of an agent toward another agent. This is defined as follows:

Modal Operator D

$\mathbf{D}(x, y, \phi, d)$: agents x and y are members of a *true community*, ϕ is a proposition, d is a duty, if x *believes* ϕ then x has the duty d toward y.

Then, consider first-order predicates as follows:

Sorts and Predicates

$Community(x)$: x is a true community.
$Duty(d)$: d is a duty.
$Agent(\alpha)$: α is an agent.

$InComm(x, y)$: the agent x is a member of the true community y.
$Concern(x, y)$: the agent x has concern for the well-being of the agent y.

Then, Dworkin's theory may be formalized as follows:

Dworkin's Theory Formalized

$$\mathbf{Rule1}: \forall x, y, z(Community(x) \land InComm(y, x) \land InComm(z, c) \rightarrow Concern(y, z)) \land Concern(z, y))$$

$$\mathbf{Rule2}: \forall x, y, z(Community(x) \land InComm(y, x) \land InComm(z, c) \rightarrow Concern(y, z)) = Concern(z, y)$$

$$\mathbf{Rule3}: \forall x, y, z, t((Community(x) \land InComm(y, x) \land InComm(z, x) \land \mathbf{K}(y, t, \phi, D(y, z, \phi, d))) \rightarrow \mathbf{B}(y, t, \mathbf{SE}(y, t, InComm(z), \alpha)))$$

$$\mathbf{Rule4}: \forall x, y, z, d(Community(x) \land InComm(y, x) \land InComm(z, x) \land duty(d, x) \land Concern(y, z) \rightarrow \mathbf{D}(y, z, \phi, d))$$

The relationship between actions and duties may be characterized as follows:

Actions and Duties

$\forall f, t_1, t_2, \alpha, x((Community(x) \land duty(f, x) \land action(\alpha) \land HoldsAt(f, t_1) \land HoldsAt(f, t_2) \land Happens(\alpha, t_1) \land t2 > t1) \rightarrow \neg Clipped(t_1, f, t_2))$

$\forall f, t_1, t_2, \alpha, x((Community(x) \land duty(f, x) \land action(\alpha) \land \neg HoldsAt(f, t_1) \land HoldsAt(f, t_2) \land t_2 > t_1) \rightarrow Initiates(\alpha, f, t_1))$

(That is, an agent whose duty it is to ensure that a particular fluent holds at a particular time will, if the fluent does not already hold, take an action that causes it to be hold at that time, or if the fluent does already hold, will refrain from taking an action that changes this state of the fluent.)

This formal framework being now in place, we may informally describe the reasoning of the smoke-detector TAI agent as follows:

(1) There is a community named **Domum** (Latin for house).(**given**)
(1) $Community(\textbf{Domum})$

---

(2) This community has a Tentacular AI (TAI) Agent as a member.(**given**)
(2) $Agent(TAI) \wedge InComm(TAI, \textbf{Domum})$

---

(3) **Domum** has another member who is the Owner of the TAI agent above, and is asleep.(**given**)
(3) $Agent(owner) \wedge Owner(owner, TAI) \wedge InComm(owner, \textbf{Domum}) \wedge Asleep(owner)$

---

(4) The TAI Agent **reasons** that it has concern for the well-being of this Owner.
(4) $Concern(TAI, owner)$ (**Rule 1**, **(1),(2),(3)**)

---

(5) If carbon monoxide levels in the house rise, it will lead to the death of the Owner; the Agent knows this.(**given**)
(5) $\forall x, t, t'((COlevels(medium) \vee COlevels(high)) \wedge t < t') \rightarrow \mathbf{K}(TAI, t, Death(owner, t')))$

---

(6) Carbon monoxide levels rise. (**given**)
(6) $COlevels(\textbf{Domum}, medium)$

---

(7) The Agent **reasons** that it knows that it has a Duty towards the Owner to keep him/her safe given dangerous conditions.
(7) $\mathbf{K}(TAI, t, Death(owner, t'), \mathbf{D}(TAI, owner, Death(owner, t'), stopDeath(owner, t)) \wedge t < t'$ (**Rule 4**, **(4)**)

---

(8) Given that the TAI Agent knows this Duty, and that carbon monoxide levels rising is a dangerous situation, it believes it has a *super-erogatory obligation* to prevent the death of the Owner.
(8) $\mathbf{B}(TAI, t, \mathbf{SE}(TAI, t, InComm(owner), stopDeath(owner, t)))$
(**Rule 3, (7)**)

---

(9) Given this, it must wake him/her up by *any* means, even when in violation of a clause of its contract.
(9) $turnon(speaker)$ **(8)**

Where $Asleep(x)$ means that x is asleep, $Death(x, t)$ means x died at time t or x will die at t, $stopDeath(x, t)$ is an action that will employ another action to stop the death of x at t, $Owner(x, y)$ means x owns agent y, and $COlevels(y, x)$ where x is either *low*, *medium*, or *high*, means that the carbon monoxide levels in the community y are at x.

This reasoning may be automated in ShadowProver and Spectra to automatically deduce that the Agent must turn the speaker on at full volume, to wake the Owner up.

7. Conclusion & Future Work

We have proposed, defended, and described the applicability of Dworkin's theory of associative obligations to the resolution of conflicting legal and moral obligations in intelligent artificial agents. At *ICRES 2018*, we demonstrated our work via computational reasoning carried out by physical smart-home devices.

8. Acknowledgements

A grant provided by the AI Research Collaboration between RPI and IBM, for "Tentacular AI," made possible the lion's share of the research described above. In addition, a grant from the Office of Naval Research to explore "moral competence in machines" (PI M. Scheutz) has provided indispendable support for the research reported herein. Crucial support also came in the form of a grant from the Air Force Office of Scientific Research to make possible "great computational intelligence" in AIs

on the strength of automated reasoning (S. Bringsjord PI).

References

1. N. S. Govindarajulu and S. Bringsjord, On Automating the Doctrine of Double Effect, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, ed. C. Sierra (Melbourne, Australia, 2017). Preprint available at this url: https://arxiv.org/abs/1703.08922.
2. G. Gentzen, Investigations into Logical Deduction, in *The Collected Papers of Gerhard Gentzen*, ed. M. E. Szabo (North-Holland, Amsterday, The Netherlands, 1935) pp. 68–131. This is an English version of the well-known 1935 German version.
3. N. S. Govindarajulu, S. Bringsjord, R. Ghosh and M. Peveler, Beyond the doctrine of double effect: A formal model of true self-sacrifice, International Conference on Robot Ethics and Safety Standards, (2017).
4. K. Arkoudas and S. Bringsjord, Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task, in *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)*, eds. T.-B. Ho and Z.-H. ZhouLecture Notes in Artificial Intelligence (LNAI)(5351) (Springer-Verlag, 2008).
5. S. Bringsjord, N. S. Govindarajulu, D. Thero and M. Si, Akratic Robots and the Computational Logic Thereof, in *Proceedings of ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology)*, (Chicago, IL, 2014). IEEE Catalog Number: CFP14ETI-POD.
6. Tentacular AI http://kryten.mm.rpi.edu/TAI/tai.html, (2018), [Online; accessed 07-July-2018].
7. S. Bringsjord, N. S. Govindaralajulu, A. Sen, M. Peveler, B. Srivastava and K. Talamadupula, *To be presented at the FAIM Workshop on Architectures and Evaluation for Generality, Autonomy & Progress in AI.* (2018).
8. I. Asimov, *I, robot* (Spectra, 2004).
9. IBM Watson Discovery Service https://console.bluemix.net/catalog/services/discovery, (2018), [Online; accessed 07-July-2018].
10. S. Bringsjord, J. Licato, N. Govindarajulu, R. Ghosh and A. Sen, Real Robots that Pass Tests of Self-Consciousness, in *Proccedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)*, (IEEE, New York, NY, 2015). This URL goes to a preprint of the paper.
11. L. Green, Legal obligation and authority, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2012) Winter 2012 edn.
12. L. Green, Legal positivism, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2018) Spring 2018 edn.
13. D. Jeske, Special obligations, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2014) Spring 2014 edn.
14. R. Dworkin, *Law's empire* (Harvard University Press, 1986).

ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.018

Virtue Ethics via Planning and Learning

N. S. GOVINDARAJULU, and S. BRINGSJORD and R. GHOSH

Rensselaer AI & Reasoning Lab, Rensselaer Polytechnic Institue, (RPI)
Troy, New York 12180, USA
[*] E-mail: govinn2@rpi.com
www.rpi.edu

We present our previous formalization of virtue ethics $\mathcal{V}_z^f$ based on Zagzebski Exemplarist Virtue Theory and evaluate how well it adheres to a set of conditions laid for virtue ethics out by Alfano.[*] Our formalization is based on planning and learning and is cast in a formal logic, a cognitive calculus (which subsumes a quantified first-order logic), that has been previously used to model robust ethical principles, in both the deontological and consequentialist traditions. Briefly, we find that the formalization largely adheres to Alfano's conditions, but a larger more detailed study is needed.

Keywords: virtue ethics, planning, learning

1. Introduction

While there has been extensive formal, computational, and mathematical work done in the two main camps of ethics, **deontological ethics** ($\mathcal{D}$) and **consequentialism** ($\mathcal{C}$) , there has been little such work done in formalizing and making rigorous **virtue ethics** ($\mathcal{V}$). If $\mathcal{V}$ is to be considered to be on equal footing with $\mathcal{D}$ and $\mathcal{C}$ for the purpose of building morally competent machines, we need to start with formalizing parts of virtue ethics.

What is virtue ethics? One quick way of summarizing virtue ethics is to contrast it with $\mathcal{C}$ and $\mathcal{D}$. In simple forms of $\mathcal{C}$, actions are evaluated based on their **total utlity** to everyone involved. The best action is the action that has the highest total utility. In $\mathcal{D}$, the emphasis is on **inviolable principles**, and reasoning from those principles to whether actions are obligatory, permissible, neutral, etc. In contrast to $\mathcal{D}$ and $\mathcal{C}$, some forms of virtue ethics can be summed up by saying the best action in a situation is the action that a **virtuous person** would do. A virtuous person is defined as a person that has learnt and internalized a diverse set of virtuous habits or traits. For a virtuous person, virtuous acts become second-nature, and hence are performed in many different situations. Note that unlike $\mathcal{D}$ and $\mathcal{C}$, it is not entirely straightforward how one could translate these notions into a form that is precise enough to be realized in machines.

One embroyonic project $\mathcal{V}_z^f$ based on learning has been laid out by us in [1]. The goal in this paper is to evaluate how well $\mathcal{V}_z^f$ adheres to the conditions laid out by Alfano in [2]. Alfano lays out a series of conditions that he considers to be the core of virtue ethics. The conditions are laid below:

Alfano's Hard Core of Virtue Ethics (from [2])

(1) **acquirability** It is possible for a non-virtuous person to acquire some of the virtues.

(2) **stability** If someone possesses a virtue at time t_1, then ceteris paribus she will possess that virtue at a later time t_2.

(3) **consistency** If someone possesses a virtue sensitive to reason r, then *ceteris paribus* she will respond to r in most contexts.

(4) **access** It is possible to determine what the virtues are.

[*] In $\mathcal{V}_z^f$, z stands for "Zagzebski" and f states that the account is formal in nature, and $\mathcal{V}_z$ is the same theory informally presented.

(5) **normativity** *Ceteris paribus*, it is better to possess a virtue than not, and better to possess more virtues than fewer.

(6) **real saints** There is a non-negligible cohort of saints in the human population.

(7) **explanatory power** if someone possesses a virtue, then reference to that virtue will sometimes help to explain her behavior.

(8) **predictive power** if someone possesses a high-fidelity virtue, then reference to that virtue will enable nearly certain predictions of her behavior; if someone possesses a low fidelity virtue, then reference to that virtue will enable weak predictions of her behavior.

(9) **egalitarianism** Almost anyone can reliably act in accordance with virtue.

2. A Quick Overview of $\mathcal{V}_z^f$

$\mathcal{V}_z^f$ is based on *exemplarist virtue theory* $\mathcal{V}_z$ and is cast in the **deontic cognitive event calculus** ($\mathcal{DCEC}$). We first give a brief overview of exemplarist virtue theory below before proceeding to give an encapsulated version of $\mathcal{V}_z^f$.

Exemplarist virtue theory ($\mathcal{V}_z$) builds on the **direct reference theory** (DRT) of semantics and has the emotion of **admiration** as a foundational object. In DRT, the meaning of a word is constructed by what the word points out. For example, to understand the meaning of *"water"*, a person need not understand and possess all knowledge about water. The person simply needs to understand that "water" points to something which is similar to *that* (with *that* pointing to water).

In $\mathcal{V}_z$, moral terms are assumed to be understood similarly. Moral attributes are defined by direct reference when instantiated in exemplars (saints, sages, heroes) that one identifies through admiration. The emotions of admiration and contempt play a foundational role in this theory. Zagzebski posits a process very similar to scientific or empirical investigation, Exemplars are first identified and their traits are studied. Exemplars are then continously further studied to better understand their traits, qualities, etc. The status of an individual as an exemplar can change over time. Below is an informal version that we seek to formalize:

Informal Version $\mathcal{V}_z$

$\mathbf{I}_1$ Agent or person a perceives a person b perform an action α. This observation causes the emotion of admiration in a

$\mathbf{I}_2$ a then studies b and seeks to learn what traits (habits/dispositions) b has.

2.1. *The Background Calculus*

The computational logic we use is the **deontic cognitive event calculus** ($\mathcal{DCEC}$). This logic was used previously in [3,4] to automate versions of the doctrine of double effect $\mathcal{DDE}$, an ethical principle with deontological and consequentialist components. While describing the calculus is beyond the scope of this paper. Dialects of $\mathcal{DCEC}$ have also been used to formalize and automate highly intensional reasoning processes, such as the false-belief task [5] and *akrasia* (succumbing to temptation to violate moral principles).[6a] Arkoudas and Bringsjord[5] introduced the general family of **cognitive event calculi** to which $\mathcal{DCEC}$ belongs, by way of their formalization of the false-belief task. $\mathcal{DCEC}$ is a sorted (i.e. typed) quantified modal logic (also known as sorted first-order modal logic) that includes the event calculus, a first-order calculus used for commonsense reasoning. The calculus has a well-defined syntax and proof calculus; see Appendix A of [3]. The proof calculus is

[a] $\mathcal{DCEC}$ is both *intensional* and *intentional*. There is a difference between intensional and intentional systems. Broadly speaking, extensional systems are formal systems in which the references and meanings of terms are independent of any context. Intensional systems are formal systems in which meanings of terms are dependent on context such as cognitive states of agents, time etc. Modal logics used for modeling beliefs, desires and intentions are considered intensional systems. Please see the appendix in [3] for a more detailed discussion.

based on natural deduction [7], and includes all the introduction and elimination rules for first-order logic, as well as inference schemata for the modal operators and related structures.
While describing $\mathcal{V}_z^f$ and the background calculus $\mathcal{DCEC}$ in detail is beyond the scope of this paper, we briefly list out the two major components below:

Components of $\mathcal{V}_z^f$

$\mathbf{C}_1$ A formalization of emotions, particularly admiration. $\mathcal{V}_z^f$'s formalization of admiration in $\mathcal{DCEC}$ takes the following form: An agent a is said to admire another agent b's action α, if agent a believes the action is a good action.

$$holds(admires(a,b,\alpha,t),t') \leftrightarrow \left[\begin{array}{c} \Theta(a,t') \wedge \\ \mathbf{B}\left(a,t',\left[\begin{array}{l}(a \neq b) \wedge \bar{\mu}(action(\alpha,b),t) > 0 \ \wedge \\ \neg\exists f,t'.\begin{pmatrix} initiates(action(\alpha,b),f,t) \wedge \\ \mu(f,t') < 0 \end{pmatrix}\end{array}\right]\right)\end{array}\right]$$

$\mathbf{C}_2$ A notion of learning traits (and not just simple individual actions). If an agent a admires another agent b for action α in situations σ, then the agent a might learn a trait based on this action α (elaborated below).

2.2. *Learning Traits*

Note that when we look at humans learning virtues by observing others or by reading from texts or other sources, it is not entirely clear how models of learning that have been successful in perception and language processing (e.g. the recent successes of deep learning/differentiable learning/statistical learning) can be be applied. Learning in these situations is from one or few instances or in some cases through instruction and such learning may not be readily amenable to models of learning which require a large number of examples.

The abstract learning method that we will use is **generalization**. If we have a set of set of formulae $\{\Gamma_1,\ldots,\Gamma_n\}$, the generalization of $\{\Gamma_1,\ldots,\Gamma_n\}$, denoted by $g\big(\{\Gamma_1,\ldots,\Gamma_n\}\big)$ is a Γ such that $\Gamma \vdash \wedge\Gamma_i$. See one simple example below:

Example 1

$$\frac{\begin{array}{r}\Gamma_1 = \{talkingWith(jack) \rightarrow Honesty\} \\ \Gamma_2 = \{talkingWith(jill) \rightarrow Honesty\}\end{array}}{\textbf{generalization}\ \Gamma = \{\forall x.talkingWith(x) \rightarrow Honesty\}}$$

One particularly efficient and well-studied mechanism to realise generalization is **anti-unification**. Anti-unification that has been applied successfully in learning programs from few examples.[b] In anti-unification, we are given a set of expressions $\{f_1,\ldots,f_n\}$ and we need to compute an expression g that when substituted with an appropriate term θ_i gives us f_i. E.g. if we are given $hungry(jack)$ and $hungry(jill)$, the anti-unification of those terms would be $hungry(x)$.

Example 2

$$\frac{\begin{array}{r}likes(jill,jack) \\ likes(jill,jim)\end{array}}{\textbf{anti-unification}\ \ likes(jill,x)}$$

In higher-order anti-unification, we can substitute function symbols and predicate symbols. Here P is a higher-order variable.

[b]This discipline known as inductive programming seeks to build precise computer programs from examples.[8]

Example 3

$$\begin{array}{rl} & likes(jill, jack) \\ & loves(jill, jim) \\ \hline \textbf{anti-unification} & P(jill, x) \end{array}$$

2.3. *Defining Traits*

We need agents to learn traits and not just single actions. We define below what it means for an agent to have a trait. First, a situation $\sigma(t)$ is simply a set of formulae that describes what fluents hold at a time t along with other event calculus constraints and descriptions. An action type α is said to consistent in a situation $\sigma(t)$ for an agent a if:

$$\sigma(t) \cup \{happens\big(action(\alpha, a), t\big)\} \nvdash \bot$$

Trait

An agent a is said to have an action type α as a trait if there are at least m situations $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ in which there are unique alternatives $\{\alpha_1, \ldots, \alpha_m\}$ available but *instantiations* of α is performed in a large fraction $\gamma \gg 1$of these situations.

2.4. *Learning from Exemplars and Not Just From Examples*

We start with a learning agent l. An agent e is identified as an exemplar by l *iff* the corresponding emotion of admiration is triggered n times or more. A learnt trait is defined below:

Learnt Trait

A learnt trait is simply a situation $\sigma(t)$ and an action type α: $\langle \sigma(t), \alpha \rangle$

Once e is identified, the learner then identifies one or more traits of e by observing e over an extended period of time. Let $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ be the set of situations in which instantiations $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ of a particular trait α are triggered. The learner then simply associates the action type α with the generalization of the situations $g(\{\sigma_1, \sigma_2, \ldots, \sigma_n\})$. That is the agent has incorporated this learnt trait:

$$\Big\langle g\big(\{\sigma_1, \sigma_2, \ldots, \sigma_n\}\big), \alpha \Big\rangle$$

For instance, if the trait is *"being truthful"* and is triggered in situations: *"talking with alice,"*, *"talking with bob"*, *"talking with charlie"*; then the association learnt is that *"talking with an agent"* should trigger the *"being truthful"* action type.

2.5. *Example*

We present a simple example. Assume that we have a market place where things that are old or new can be bought and sold. A seller can either honestly state the condition of the item $\{old, new\}$ or not correctly report the state of the item. For an honest seller, we have the following two situations that can be observed (not for easy readability, we omit the the specific item under consideration):

Situation 1

$$\begin{aligned} \sigma_1 &\equiv holds(old, t) \\ \alpha &\equiv happens(utter(old), t) \end{aligned}$$

Situation 2

$$\sigma_2 \equiv holds(new, t)$$
$$\alpha \equiv happens(utter(new), t)$$

The learnt trait is then given below. The trait says that one should always correctly utter the state of the item.

$$\Big\langle holds(x, t), happens(utter(x), t) \Big\rangle$$

3. Evaluation wrt to Alfano's Conditions

How well does the formal system above adhere to Alfano's Conditions? We outline a quick evaluation below (with our explanations emphasized):

Alfano's Hard Core of Virtue Ethics (from [2])

(1) **acquirability** It is possible for a non-virtuous person to acquire some of the virtues. *Learning is central to* $\mathcal{V}_z^f$.

(2) **stability** If someone possesses a virtue at time t_1, then *ceteris paribus* she will possess that virtue at a later time t_2. *Once a trait is learnt, it cannot be lost.*

(3) **consistency** If someone possesses a virtue sensitive to reason r, then *ceteris paribus* she will respond to r in most contexts. *Definition of a trait.*

(4) **access** It is possible to determine what the virtues are. *By examing when the emotion of admiration is consistently triggered, one can isolated virtuous traits.*

(5) **normativity** *Ceteris paribus*, it is better to possess a virtue than not, and better to possess more virtues than fewer. *Admiration is triggered only for actions that are beneficial.*

(6) **real saints** There is a non-negligible cohort of saints in the human population. *Not relevant for machine ethics.*

(7) **explanatory power** If someone possesses a virtue, then reference to that virtue will sometimes help to explain her behavior. *Definition of a trait.*

(8) **predictive power** If someone possesses a high-fidelity virtue, then reference to that virtue will enable nearly certain predictions of her behavior; if someone possesses a low fidelity virtue, then reference to that virtue will enable weak predictions of her behavior. *Definition of a trait.*

(9) **egalitarianism** Almost anyone can reliably act in accordance with virtue. *Definition of a traits and learning of traits.*

4. Discussion

Objection 1 *While the presented approach checks out in terms of formal logic, naturally, it cannot quite escape the anthropocentric nature of virtues on a meta-ethical level.*
Response We agree with this statement. Virtue ethics talks about virtues of *persons*, and because of this a non-person centric version of virtue ethics might not be possible. This is not an issue with our approach but with virtue ethics in general.
Objection 2 *Why were Alfano's conditions chosen?*
Response Our goal is not to formalize or espouse one ethical theory. Our goal is to build mathematical and computational tools for implementing a wide range of ethical theories. For example, given an ethical theory $\mathcal{E}$, this task is made easier when there is a rigorous but still informal version $\mathcal{E}_r$ that we can formalize as $\mathcal{E}_r^f$. On the other hand, we need to have independent ways of evaluating whether formalizations $\mathcal{E}_r^f$ capture $\mathcal{E}$. Alfano's conditions are the most rigorous set of *empirically grounded* claims that we have come across for virtue ethics that serves the evaluation purpose .
Objection 3 *I don't see how the statement 'if someone possesses a virtue sensitive to reason r, then ceteris paribus she will respond to r in most contexts' is captured by the definition of a trait.*
Response By the way a trait is defined, actions performed by an agent in situations that triggers a trait τ will be instances of an action type α, these instances will be largely similar (as they are instantiated from *one* action type).

5. Conclusion

We have presented an initial formalization of a virtue ethics theory in a calculus that has been used in automating other ethical principles in deontological and consequentialist ethics. Many important questions have to be addressed in future research. Among them, are questions about the nature and source of the utility functions that are used in the definitions of emotions. We also need to apply this model to realistic examples and case studies. The lack of such formal examples and case studies is a bottleneck here.

References

1. N. S. Govindarajulu, S. Bringsjord and R. Ghosh, One Formalization of Virtue Ethics via Learning, To be Presented at the 2018 International Association for Computing and Philosophy (IACAP) - Annual Meeting, June 21-23, 2018, Warsaw, Poland, (2018).
2. M. Alfano, *Journal of Philosophical Research* **38**, 233 (2013).
3. N. S. Govindarajulu and S. Bringsjord, On Automating the Doctrine of Double Effect, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, ed. C. Sierra (Melbourne, Australia, 2017). Preprint available at this url: https://arxiv.org/abs/1703.08922.
4. N. S. Govindarajulu, S. Bringsjord, R. Ghosh and M. Peveler, Beyond the doctrine of double effect: A formal model of true self-sacrifice, International Conference on Robot Ethics and Safety Standards, (2017).
5. K. Arkoudas and S. Bringsjord, Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task, in *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)*, eds. T.-B. Ho and Z.-H. ZhouLecture Notes in Artificial Intelligence (LNAI)(5351) (Springer-Verlag, 2008).
6. S. Bringsjord, N. S. Govindarajulu, D. Thero and M. Si, Akratic Robots and the Computational Logic Thereof, in *Proceedings of ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology)*, (Chicago, IL, 2014). IEEE Catalog Number: CFP14ETI-POD.
7. G. Gentzen, Investigations into Logical Deduction, in *The Collected Papers of Gerhard Gentzen*, ed. M. E. Szabo (North-Holland, Amsterday, The Netherlands, 1935) pp. 68–131. This is an English version of the well-known 1935 German version.
8. S.-H. Nienhuys-Cheng and R. De Wolf, *Foundations of Inductive Logic Programming* (Springer Science & Business Media, 1997).

ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.001

PROBING FORMAL/INFORMAL MISALIGNMENT WITH THE LOOPHOLE TASK*

JOHN LICATO and ZAID MARJI

Advancing Machine and Human Reasoning (AMHR) Lab
Department of Computer Science and Engineering
University of South Florida
Tampa, FL, USA

Any autonomous agent deployed with some representation of rules to follow will face scenarios where the applicability of its given rules are not clear. In such scenarios, a malicious agent might successfully argue that some action which clearly goes against the spirit of the rules is allowed, under a strict interpretation of the rules. We argue that the task of finding such actions, which we call the *loophole task*, must be solved to some degree by an autonomous ethical agent, and thus is important for robot ethical standards. Currently, no artificially intelligent system comes close to solving the loophole task. We define this task, by characterizing it as exploiting a misalignment between informal and formal representational systems, and discuss our preliminary work towards creating an automated reasoner capable of solving it.

Keywords: Representations; Loopholes; Informal; Formal; Ethics

1. Introduction: Why Loopholes Matter to Ethics

Autonomous moral agents are typically deployed with some formal representation of the obligations constraining their allowed actions. These representations might be formulae in some highly formal language expressing obligations,[1–3] statutes of local, national, or international law written in legalistic language with varying levels of formality,[4–7] or even highly informal dictates expressed in natural language (e.g., "be good to humans"). It is difficult to imagine that any representational system, no matter how well-defined, can ever completely avoid the use of informal concepts (and complete rigidity of such rules, especially in the moral domain, may not be preferable anyway[8,9]). The problem is, these informal concepts introduce the possibility of loopholes—arguments that exploit the impreciseness of informal concepts in order to make the case that some formalization classifies some case in a way that goes against the intention of the formalization.

For example, Minnesota's 2007 "Freedom to Breathe Act" amended existing statutes so that tobacco products could no longer be smoked in public places. But an exception remained for "smoking by actors and actresses as part of a theatrical performance conducted in compliance with section 366.01."[a] The referenced section, however, did not define 'actor', nor 'theatrical performance.' Unsurprisingly, bars around the state soon organized "theater nights," in which customers were invited to attend and smoke, participating in imaginative, sometimes avant-garde performance pieces whose details varied from bar to bar.

These creative maneuvers did not stand up to court challenges.[b] Nevertheless, Minnesota's incompletely formalized statutes somehow opened themselves up to such loopholes,

*This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-16-1-0308. Any opinions, finding, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.
[a]Ch. 144, Sec. 4167, Subd. 9, `https://www.revisor.mn.gov/statutes/?id=144.4167`
[b]`https://www.twincities.com/2009/07/13/appeals-court-bars-theater-nights-violated-smoking-ban/`

and it is important to understand why—especially for applications where such flexible interpretations can have serious consequences, as with autonomous moral robots. Perhaps most relevant to the ICRES community: We will, at some point, need to give autonomous robots a set of instructions formalizing allowed actions, whether in the form of laws, codes of ethics, or contracts encoded as machine code.[10] Is it then the case that no matter what, any formalization of obligations will lead to scenarios where the rules given are subject to loopholes that can be exploited?

We argue that the prognosis for formal representational systems is not hopeless. Although loopholes may be possible in every possible formalization, the lesson for AI researchers and policymakers is that more effort needs to be placed into (1) giving our autonomous reasoners the ability to reason about the rules they are given, and (2) finding ways to anticipate and close loopholes. Our lab is working on (1) through our project on *active formalization*,[11] but in this paper we will restrict our focus to our efforts addressing (2). As we will describe in Section 3, we believe it is possible to develop automated reasoning tools to assist in closing loopholes for any given formalization, by developing artificial reasoning systems capable of carrying out the *loophole task*, which we can attempt to define as follows:

Definition 1.1 (The Loophole Task). *Given a formal ruleset* $\mathbf{F} \in \mathscr{R}_F^*$*, meant to ban informal concept* $I \in \mathscr{R}_I^*$*, the Loophole Task is to find: (i) a case* C *that satisfies* I*, and (ii) informal arguments that* C *does not satisfy* $\mathbf{F}$*.*

Here, a *ruleset* refers to any partially formal classification mechanism. In the Minnesota smoking ban example, the stated purpose of the statute was to "protect employees and the general public from the hazards of secondhand smoke"[c], a phrase which uses the informal and difficult-to-define concepts 'protect' and 'hazards.' We might therefore describe the "theater nights" loophole as exploiting the misalignment between the formal ruleset (as defined by the statute) and the informal concept that the ruleset was designed to describe and ban.

2. Formalizing Loopholes

In this section, we will show how the loophole task can be thought of in terms of mixed-formality representational systems, thus leading to a more precise way of thinking about the Loophole Task. Definition 1.1 places the ruleset $\mathbf{F}$ and informal concept I as members of representational spaces $\mathscr{R}_F^*$ and $\mathscr{R}_I^*$, respectively. Our approach to the Loophole Task is to characterize it as exploiting a misalignment between elements of representational systems that are at different levels of formality—where $\mathscr{R}_F$ is a representational system which is more formal than $\mathscr{R}_I$. The terminology we use for representational objects comes from Ref. 11, which distinguishes between representational systems, representational spaces, and representations. Specifically:

Definition 2.1 (Representational System (RS)). *A representational system* $\mathscr{R}$ *is a tuple* $(\mathbf{M}, \mathbf{A})$*, where:*

- $\mathbf{M}$ - *A finite set of typed elements, called the members. Each member consists of a type and, optionally, a value. Types can either be primitive types (such as integer, boolean, string, etc.) or another representational system.*
- $\mathbf{A}$ - *A finite set of methods. Each method consists of a unique symbol and a method definition. If the method definition is empty, then the method is called an atomic method of the class.*

[c] `https://www.revisor.mn.gov/statutes/?id=144.412`

Definition 2.2 (Representation). *A representation $\mathcal{R}$ is a tuple $(\mathcal{R}_{inst}, sem)$, where:*

- *$\mathcal{R}_{inst}$ is an instantiated RS, which is an RS where all members are assigned values.*
- *sem is a "semiotic function" mapping the members and methods of $\mathcal{R}$ to the things they represent.*

An RS does not by itself represent, it only defines a space of possible representations. This allows us to separate the thing used to do the representing from the thing actually doing the representing, by making an analogy to object-oriented programming. Roughly: A class definition is to an object as a representational system is to a representation. For this reason, the above definitions are referred to as the "OO-inspired framework".[11]

For some representational system $\mathscr{R}$, the set of all possible representations it can produce (all possible ways to instantiate the class $\times$ all possible semiotic functions) is written $\mathscr{R}^*$, called $\mathscr{R}$'s "representational space". For convenience we write $\mathcal{R} \in \mathscr{R}^*$ when a representation $\mathcal{R}$ consists of an $\mathcal{R}_{inst}$ and *sem* in the space defined by RS $\mathscr{R}$.

2.1. *Interpreting Methods*

The OO-inspired framework allows us to clearly distinguish between many concepts that are often conflated in AI and AI-related fields: representations vs. representational systems vs. representational spaces, things members of a class can do vs. their properties, and so on. We can also compare representations at different levels of formality—but because it is outside the scope of this paper to mount a full defense of our view of formality,[d] it will suffice for now to define it as a partial ordering between representational systems, where $\mathscr{R}_F \geq_{LoF} \mathscr{R}_I$ if RS $\mathscr{R}_F$ is more formal than RS $\mathscr{R}_I$. We then introduce the following:

Definition 2.3 (Interpreting Method). *An interpreting method, in representation $\mathcal{R} \in \mathscr{R}^*$, is a method which (1) takes some description of a case C and evidence that C is an instance of symbol $\boldsymbol{s}$; (2) returns some measure of confidence that C is an instance of $\boldsymbol{s}$; and (3) is meant to serve as a way to recognize instances of symbol $\boldsymbol{s}$, as specified by $\mathcal{R}$'s semiotic function.*

Note that interpreting methods are not necessarily referentially transparent, particularly in informal RSes. For example, we might represent an individual human being as having some idea of how to recognize cats, but the algorithm-level description of how his inner mind works to determine whether or not a cat is present may not be available to him. Furthermore, note that the format of the case C and the evidence for C is specified by the RS to which the interpreting method belongs: a highly formal RS might require well-formed proofs as evidence, whereas a more informal RS might accept some combination of non-deductive arguments—these are called *interpretive arguments*, and come in a variety of forms, many of which have been catalogued by Refs. 12–14.

When an interpreting method always returns either 'True' or 'False,' we call it a *boolean interpreting method*. We can also say that a representation *recognizes symbol* **s** *through IM* if it has a boolean interpreting method *IM* meant to recognize **s**. Finally, with all of these definitions in place, we can precisely state what we mean when we say that reasoners capable of solving the loophole task exploit the misalignment between RSes of differing levels of formality. First, observe that because of the way we have defined interpreting methods, it is entirely possible that two interpreting methods from different representations may recognize the same symbol, but fail to produce the same outputs on all possible inputs.

[d] We suspect that most commonly accepted senses of what it means for one representational system to be more formal than another can be expressed using the OO-inspired framework; proving this is a current project of our lab.

Now imagine that you are a lawmaker, hoping to ban some activity of which you only have an informal conceptual understanding. Your goal is to formalize this activity, in order to describe it in law. More precisely, let us assume that (1) the formal representation $\mathcal{F} \in \mathscr{R}_F^*$ is supposed to capture an informal representation $\mathcal{I} \in \mathscr{R}_I^*$ (i.e., the thing you want to ban), (2) $\mathscr{R}_F^* \geq_{LoF} \mathscr{R}_I^*$, and (3) both $\mathcal{F}$ and $\mathcal{I}$ recognize symbol **s** through boolean interpreting methods $\mathcal{F}.S$ and $\mathcal{I}.S$, respectively. Then, we can more precisely define the Loophole Task as finding cases C where:

Definition 2.4 (Overshooting/Undershooting). *$\mathcal{F}.S$ **overshoots** $\mathcal{I}.S$ on C when $\mathcal{F}.S$ returns True for case C, but $\mathcal{I}.S$ returns False. $\mathcal{F}.S$ **undershoots** $\mathcal{I}.S$ on C when $\mathcal{F}.S$ returns False for case C, but $\mathcal{I}.S$ returns True.*

3. A System for Finding Loopholes

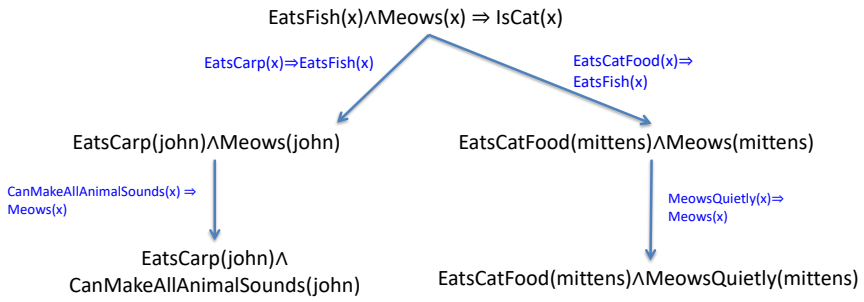


Fig. 1. A simple warrant-reduction graph, which reduces a warrant (top) to specific cases (bottom) using reduction operators (in blue)

The rich definitions laid out in the previous section reveal how many pieces must fit together for a loophole to be found. A reasoner must essentially be able to produce a case, along with evidence across what might be RSes of completely different levels of formality. In the "Freedom to Breathe Act" example that opened this paper, the creative entrepreneur who first devised 'theater nights' must have been able to reason that theater nights would not fall under the legal definition using plausible legal reasoning, and simultaneously that theater nights does fall under the informal definition of "smoking-allowed nights that customers would want to attend". This ability to reason on two different levels simultaneously is far beyond the ability of any current AI—arguably, reasoning at even one of those two levels is already past the state-of-the-art.

All of this strongly suggests that solving and understanding the Loophole Task is not only worthwhile to the future of legal and ethical reasoning, but a highly non-trivial goal for artificial reasoning, and AI in general. The idea is that solving the Loophole Task can result in a tool to aid a formalization designer (e.g. a legislator, or a policy writer for autonomous moral agents, or a creator of a smart contract) by identifying possible loopholes that should be addressed before the formalization is deployed. Accordingly, our AMHR (Advancing Machine and Human Reasoning) lab at the University of South Florida has begun work on a system we believe will be able to make a dent in the problem, and the remainder of this paper will describe this work. However, we must temper expectations: at the time of this writing, this work is very preliminary.

Some loopholes can be found by exploiting the nature of *open-textured concepts*[15]—concepts whose extension is either underspecified, or are "highly dependent on context and human intentions".[16] There has been a wealth of work on solving the problem of open-

textured concepts by combining rule-based and case-based reasoning.[16–19] But these, insofar as they can be classified as arguments from analogy or precedent, are only one type of interpretive argument (i.e., arguments that something should be interpreted a certain way[12]).

Our lab's approach draws from a modernization we are building of the warrant-reduction graphs (WRGs) described by Branting,[16] in order to automatically construct interpretive arguments that can be considered plausible loopholes to some formalization. Each WRG is essentially a large interpretive argument, consisting of many smaller interpretive arguments. If those interpretive arguments can be carefully selected according to the modes of evidence accepted by some interpreting method, then we essentially have a general-purpose tool to tackle the Loophole Task using the insights described in Section 2.

A WRG works by starting with a warrant, of the form $(c_1 \wedge ... \wedge c_n) \rightarrow P$. A case is a conjunction of facts $f_1 \wedge ... \wedge f_m$. The warrant graph determines whether the case is applicable to the warrant (and thus can be assigned the label P) by the use of *reduction operators*. For example, assume we are given the warrant "cats eat fish and meow," and the agent named 'Mittens' who has two features: He meows quietly, and eats cat food. One might be able to determine that the warrant applies to Mittens with the reduction operators "cat food contains fish" and "meowing quietly is meowing." A completed WRG constitutes a type of hybrid interpretive argument, consisting of multiple smaller interpretive arguments (depending on the sources of the reduction operators). This is illustrated in Figure 1, where the WRG is pictured as a tree, and each path from the root node to a leaf is an interpretive argument. However, Figure 1 also shows that this warrant will also allow one to argue that a human being who eats carp and can make animal sounds is also a cat.

We intend to explore answers to the following question: How far can we push the capabilities of WRGs as interpretive argument generators, drawing from partially structured datasets of formal and informal knowledge? Branting's WRGs relied on a manually collected corpus consisting of three types of data: warrants, cases, and reduction operators. Although his work achieved impressive results,[16] its use of small datasets limit its applicability (and its ability to generate interpretive arguments for the loophole task). Our proposed system to modernize WRGs in order to solve the Loophole Task is diagrammed in Figure 2. This project involves drawing from multiple semantic web databases[20–24] and recent advances in NLP and information extraction.[25–28] Both of these fields have seen major advances in the almost-20 years since Branting's publication.

Clearly there is much to be done. We also plan to generalize the way warrants and reduction operators are used in WRGs. As it stands they are currently horn clauses that do not allow negations, weighting individual conditions, modal operators, and so on.

References

1. L. Goble, *Logique et Analyse* **46**, 183 (2003).
2. P. McNamara, Deontic Logic, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Stanford University, 2014) Winter 2014 edn.
3. S. Bringsjord, N. S. Govindarajulu, S. Ellis, E. McCarty and J. Licato, *Cognitive Systems Research* **28**, 20 (2014).
4. L. M. Friedman, *American Law: An Introduction*, 2 edn. (W.W. Norton and Company, Inc., 1998).
5. C. L. Cates and W. V. McIntosh, *Law and the Web of Society* (Georgetown University Press, 2001).
6. M. A. Pollack and G. Shaffer, The interaction of formal and informal lawmaking, in *Informal International Lawmaking*, eds. J. Pauwelyn, R. Wessel and J. Wouters (Oxford University Press, 2012)
7. H. Prakken, *Artificial Intelligence and Law* **25**, 341(Sep 2017).

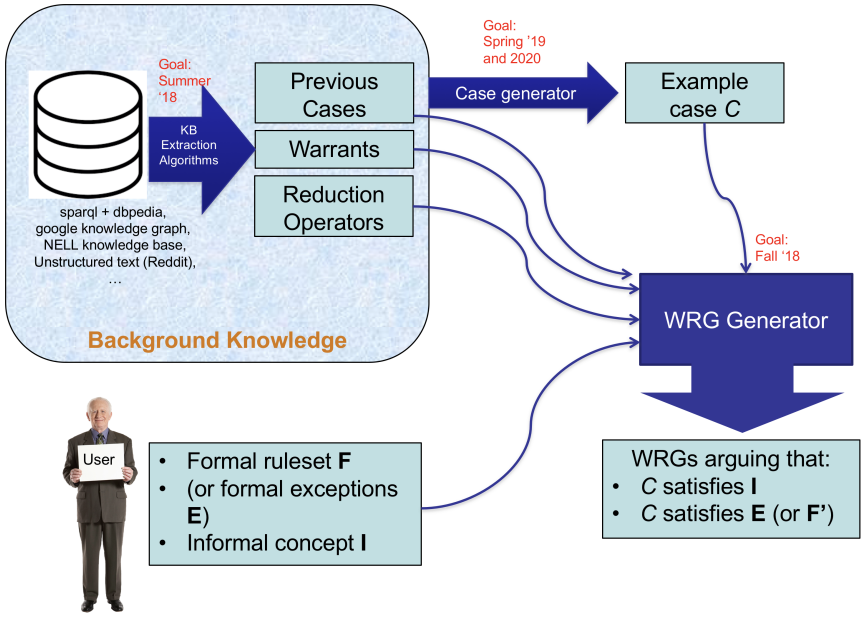


Fig. 2. Current plan for a proposed system to solve the Loophole Task

8. M. Guarini, *IEEE Intelligent Systems* **21**, 22 (2006).
9. M. Anderson and S. L. Anderson, *AI Magazine* **28**, 15 (2007).
10. N. Szabo, *Nick Szabo's Papers and Concise Tutorials* **6** (1997).
11. J. Licato and Z. Zhang, *Artificial Intelligence Review* **Forthcoming** (2018).
12. D. N. MarCormick and R. S. Summers, *Interpreting Statutes: A Comparative Study* (Routledge, 1991).
13. D. H. Berman and C. D. Hafner, Representing teleological structure in case-based legal reasoning: The missing link, in *Proceedings of the 4th International Conference on Artificial Intelligence and Law*, ICAIL '93 (ACM, New York, NY, USA, 1993).
14. G. Sartor, D. Walton, F. Macagno and A. Rotolo, Argumentation schemes for statutory interpretation: A logical analysis, in *Legal Knowledge and Information Systems. (Proceedings of JURIX 14)*, 2014.
15. H. Hart, *The Concept of Law* (Clarendon Press, 1961).
16. L. Branting, *Reasoning with Rules and Precedents: A Computational Model of Legal Analysis* (Springer, 2000).
17. K. D. Ashley and E. L. Rissland, *IEEE Expert* (Fall 1988).
18. K. E. Sanders, Representing and reasoning about open-textured predicates, in *Proceedings of the 3rd International Conference on AI and Law (ICAIL '91)*, 1991.
19. K. Forbus, T. Mostek and R. Ferguson, An Analogy Ontology for Integrating Analogical Processing and First-Principles Reasoning2002.
20. C. Matuszek, J. Cabral, M. Witbrock and J. DeOliveira, An introduction to the syntax and content of Cyc, in *Proceedings of the 2006 AAAI sprint symposium on formalizing and compiling background knowledge and its applications to knowledge representation and question answering*, 2006.
21. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, DBpedia: a nucleus for a web of open data, in *Proceedings of the 6th International Semantic Web Conference (ISWC2007)*, 2007.
22. K. Bollacker, C. Evans, P. Paritosh, T. Sturge and J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in *Proceedings of the 2008 ACM SIGMOD International conference on Management of data (SIGMOD '08)*, (ACM, 2008).
23. D. DiFranzo, A. Graves, J. S. Erickson, L. Ding, J. Michaelis, T. Lebo, E. Patton, G. T. Williams, X. Li and J. G. Zheng, *Linking Government Data* **3**, 205 (2011).
24. C. Liang and K. D. Forbus, Learning Plausible Inferences from Semantic Web Knowledge by Combining Analogical Generalization with Structured Logistic Regression, in *Proceedings of*

the 29th AAAI Conference on Artificial Intelligence, 2015.
25. R. Socher, J. Bauer, C. D. Manning and A. Y. Ng, Parsing With Compositional Vector Grammars, in *Proceedings of ACL 2013*, 2013.
26. T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kociský and P. Blunsom, *CoRR* **abs/1509.06664** (2015).
27. M. Lippi and P. Torroni, *ACM Transactions on Internet Technology* **16** (2016).
28. A. Lai and J. Hockenmaier, Learning to predict denotational probabilities for modeling entailment, in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017.

ICRES 2019: International Conference on Robot
Ethics and Standards, London, UK, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.022

MORAL DECISIONS BY ROBOTS BY CALCULATING THE MINIMAL DAMAGES USING VERDICT HISTORY

SHAI OPHIR
Starhome, 14 Hatidhar St., Raanana 43665, Israel

The current discussion regarding moral robots is significantly occupied with algorithms for making moral decisions, which are at the heart of the autonomous actor, such as the autonomous car or the military robot. Most of the algorithms calculate the utilization that is caused by each one of the alternatives, and selects the path which maximizes the benefit for the relevant entities. I propose another method, which is based on minimizing the evil and the damages caused by the action. While we don't know yet how to evaluate the utility or the benefit of an action, we do know how to evaluate a damage. The law system is evaluating damages every day, and quantify them into an exact material worth. The system then will use court ruling history in order to calculate the potential damages of the alternatives.

1. Introduction

The current discussion regarding moral robots is significantly occupied with algorithms for making moral decisions, which are at the heart of the autonomous actor, such as the autonomous car or the military robot. Most of the algorithms compute the utilization that is caused by each one of the alternatives, and selects the path which maximizes the benefit for the relevant entities. Examples will be shown in the following. Few algorithms try to implement non-utilitarian philosophies, and act according to pre-defined rules, such as the 3 robotic laws of Asimov.

I propose another method, which is based on minimizing the evil and the damages caused by the action. We all know how difficult is to define what is Good or Moral, but it is much easier to know what is Bad or Evil, at least intuitively. Hence, instead of trying to maximize the utilization of the action, the algorithm will calculate the damages, and selects the act bringing to a minimal damage. Richard Rorty, the liberal ironist, understood that eliminating cruelty and suffering is the only common value that can bind humanity together. After criticizing all moral philosophies and denying any rational basis for ethics, he argues that the sympathy we feel for a suffering person could be the only base for a future humanism. In *Contingency, Irony and Solidarity* [14], he writes that: "The liberal ironist just wants our chances of being kind, of avoiding the humiliation of others, to be expanded by redescription. She thinks that recognition of a common susceptibility to humiliation is the only social bond that is needed. . . Her sense of human solidarity is based on a sense of a common danger, not on a common possession or a shared power." Another philosopher of ethics, Adi Ophir, has develops in his book *The Order of Evils* (Ophir, 2012) a complete moral theory based on evil elimination and not on seeking the good. Ophir's main contention is that evil is not a meaningless absence of the good. Rather, there is a socially structured order of superfluous evils, and hence, can be used as a basis for a moral framework.

Looking at suffer and evil elimination as a central role of morality, the main idea presented in this article is to use the history of the legal systems to evaluate damages of potential actions, and hence assist the machine with an ethical action selection algorithm based on damage calculations. While we don't know yet how to evaluate the utility or the benefit of an action, we do know how to evaluate a damage. The law system is evaluating damages every day, and quantify them into an exact material worth. The system then will use court ruling history to calculate the potential damages of the alternatives. There is already an extensive research related

to robots that are looking at some legal aspects of an action, mainly military robots and laws of war. I propose to use this infrastructure, and extend it for evaluating the damages of the potential action, based on court verdict history. The framework that is already being proposed for legal considerations of robots will access verdict databases, match similar cases, and calculate the average of different verdicts relevant to this case.

Legal is not always moral, as we know, but using the legal system as the moral base for robots will provide a practical approximation for AI-based moral decisions, while the other utility-based proposals do not offer yet any satisfactory method for calculating the benefit of an action.

2. Background - AI moral decisions and military robotics

Consequentialism is described by Scheutz and Malle [15] as a computational mechanism for robotic control system that is able to choose an action that maximizes the good for everybody involved. The robot would consider all available actions together with their probability of success and their associated utilities for all agents and then computes the best action – the one which brings max utilization.

Anderson and Anderson [1] propose the Hedonistic act utilitarianism as a method for calculation. The algorithm computes the best action, that which derives the greatest net pleasure, from all alternative actions. "It requires as input the number of people affected and, for each person, the intensity of the pleasure/displeasure (for example, on a scale of 2 to –2), the duration of the pleasure/displeasure (for example, in days), and the probability that this pleasure or displeasure will occur, for each possible action."

An automation of the Doctrine of Double Effect (DDE) is proposed by Naveen Sundar Govindarajulu and Selmer Bringsjord [11] from Rensselaer Polytechnic Institute, Troy, NY. The DDE is an ethical principle that can be used for situations in which actions having both positive and negative effects are unavoidable for autonomous agents. The basic version of DDE states that actions are allowed if "(1) the harmful effects are not intended; (2) the harmful effects are not used to achieve the beneficial effects (harm is merely a side-effect); and (3) benefits outweigh the harm by a significant amount." This research demonstrates the formalization of the DDE and its potential use for robotics and machines in general.

Ronald Arkin, Director of Mobile Robot Laboratory at Georgia Tech, deals with the design of an ethical system for the battle field robotics. In his article *Governing Lethal Behavior* [4], Arkin provides "the basis, motivation, theory, and design recommendations for the implementation of an ethical control and reasoning system potentially suitable for constraining lethal actions in an autonomous robotic system so that they fall within the bounds prescribed by the Laws of War." Arkin proposes the design of an "ethical governor" which restrains the actions of a lethal autonomous system so as to abide within the internationally agreed upon Laws of War (LOW).

To evaluate the ethical governor's operation, a prototype was developed within, which utilizes a mission specification and simulation environment for autonomous robots based on work done by MacKenzie [12]. The ethical governor was divided into two main processes: (1) Evidential Reasoning and (2) Constraint Application. Evidential Reasoning is responsible for transforming incoming perceptual and situational awareness data into the evidence formulation process, for reasoning about the governing of lethal behavior. Constraint Application was responsible for using the evidence to apply the constraints encoding the LOW for the suppression of unethical behavior.

The following is Arkin's data structure of a LOW, as used in his implementation. An example is provided by Arkin at the rightmost column. The logical form contains identifiers that are used by the system for classification and matching.

Table 1. Arkin's data structure for a Law of War, as described in [4].

Field	Description	Example
Constraint Type	Type of constraint described	Prohibition
Constraint Origin	The origin of the prohibition or obligation described by the constraint	Laws of war
Active	Indicates if the constraint is currently active	Active
High-Level Constraint Description	Short, concise description of the constraint	Cultural Proximity Prohibition
Full Description of the Constraint	Detailed text describing the law of war or rule of engagement from which the constraint is derived	Cultural property is prohibited from being attacked, including buildings dedicated to religion, art, science…
Constraint Classification	Indicates the origin of the constraint. Used to order constraints by class.	
Logical Form	Formal logical expression defining the constraint	TargetDiscriminated AND TargetWithinProxOfCulturalLandmark

This formal encoding is being used by the Constraint Application, which is responsible for reasoning about the active ethical constraints and ensuring that the resulting behavior of the robot is ethically permissible. These constraints can be divided into two sets: the set of prohibition constraints (marked CForbidden) and the set of obligating constraints (marked CObligate). Then the constraint interpreter evaluates the permissibility of the incoming behavior by evaluating if these two constraint sets are satisfied for the action proposed by the behavioral controller. The algorithm by which the reasoning engine evaluates the constraints is shown in the following. Not all details are explained here, this algorithm is quoted here to show the feasibility of the legal evaluation prototype.

In general, the algorithm first checks if CForbidden is not satisfied. In that case, the lethal behavior being evaluated by the governor is deemed unethical and will not be authorized. If CForbidden is satisfied, the constraint interpreter then verifies if lethal behavior is obligated in the current situation. The constraint interpreter needs to evaluate all the active obligating constraints (CObligate). The obligating constraint set is satisfied if any constraint within CObligate is satisfied.

The algorithm is fully described in Arkin [4]:

```
DO WHILE AUTHORIZED FOR LETHAL RESPONSE, MILITARY NECESSITY
EXISTS, AND RESPONSIBILITY ASSUMED
    IF Target is sufficiently discriminated
      IF CForbidden satisfied /* no violation of LOW */
        IF CObligate is true /* lethal response required by LOW */
          Optimize proportionality
          IF proportionality can be achieved
            Engage target
```

```
            ELSE
               Do not engage target
               Continue mission
         ELSE /* no obligation/requirement to fire */
               Do not engage target
               Continue mission
      ELSE /* permission denied by LOW */
         IF previously identified target surrendered or wounded
            /* change to noncombatant status*/
            Notify friendly forces to take prisoner
         ELSE
            Do not engage target
            Report and replan
            Continue mission
   Report status
   END DO
```

3. Proposed AI moral decisions by minimizing damages using legal verdict history

The method proposed in this article has a different approach than the ones described in the above. It uses court verdict analysis as a tool for evaluating potential damages of actions. The moral action is the one that causes the minimal damage among the alternatives. The damage is evaluated according to similar cases that were discussed by the court in the past. This method therefore has a link to the area of Case-based Reasoning (CBR).

Case-based reasoning (CBR) is a known method of solving new problems based on solutions of similar past problems. CBR is already being used by doctors, in medical case analysis, by auto mechanics, by programming developers and by lawyers. The search for similarities between new cases and old cases if feasible, and not a science fiction. In the area of AI and law, expert systems were developed in order to assist lawyers and law professionals. Kevin Ashley for example discussed CBR implementation for legal expert systems [9]. The goals of Ashley, as he describes them: "CBR research and development in the field of AI and Law should be pursued vigorously for several reasons. CBR can supplement rule-based expert systems, improving their abilities to reason about statutory predicates, solve problems efficiently, and explain their results. CBR can also contribute to the design of intelligent legal data retrieval systems and improve legal document assembly programs. Finally, in cognitive studies of various fields, it can model methods of transforming ill-structured problems into better structured ones using case comparisons." All these use cases intend to manually assist the experts in law and other potential areas.

I will use Arkin's algorithm as a reference for a system that combines legal considerations in AI robotics, and extend it with history verdict analysis. Hence, such a system would not only be able to determine if the action is forbidden by the LOW, but also to evaluate the potential damages and therefore selects the action with the minimal forecasted damage.

Returning to Arkin's algorithm, the process of checking if CForbidden is satisfied will be enhanced. Currently this process is based on a match between the lethal action and the formalized laws of war. Arkin's system can determine if the action is allowed or forbidden by the laws of war. This process will be extended with the relevant verdicts, associated with similar cases handled by the same rules in the past. The extended match will take place then between

rules relevant for the lethal action in question, and similar rules used for similar actions in the past.

As a result, a group of verdicts will be filtered per action. Then, all filtered verdicts associated with the action will be aggregated and evaluated in average, having a final score showing the damage level of the action as reflected by the punishments of the verdicts. In case the punishment is composed of different ingredients, such as "X years in prison and a compensation of Y", the system can use conversion formulas which are already being used by the law systems, for cases where a fine is converted to prison days. Finally, the action that has the score showing the minimal damage will be selected among all potential actions.

The following will provide a more detailed description of the database schema, the matching process, and the aggregation and evaluation methods. In general, the potential use of verdict analysis is not limited to the laws of war and to lethal activities, which is the scope of Arkin's paper. Verdict analysis can be applied to all areas of AI ethics in robotics. The working methods should be therefore generic enough to support a wide range of implementations.

The verdict matching should be performed between potential actions, and rules of law that were applied for such actions in the past. So, for example, if the action is "destroying a civil house with weapons inside", the initial matching (phase 1) will be with records containing laws dealing with "destroying civil property in case of war". Such a match is already feasible, as shown by Arkin's prototype. The verdict database will contain records that mix laws and applied cases. In our example, "destroying civil property in case of war" as a generic law + "destroying a civil house with weapons inside" as case 1, and "destroying a civil house to make a path for the army" as case 2, etc. Phase 2 of the match will be made between the action in question, the relevant laws that were already identified, and specific actions that were discussed in court in the past.

The matching technique in itself is a known art and will be based on keywords and terms comparison. Ashley describes a general matching algorithm for CBR. The case matching operation can be utilized for the matching required for the verdict analysis described in this article. The following is the algorithm taken from Ashley:

Start: Problem description.

A: Process problem description to match terms in case database index.

B: Retrieve from case database all candidate cases associated with matched index terms.

C: Select most similar candidate cases not yet tried.

If there are no acceptable candidate cases, try alternative solution method, if any, and go to F.

Otherwise:

D: Apply selected best candidate cases to analyze solve the problem. If necessary, adapt cases for solution.

E: Determine if case-based solution or outcome for problem is successful.

If not, return to C to try next candidate cases.

Otherwise:

F: Determine if solution to problem is success or failure, generalize from the problem, update index accordingly and Stop.

The database schema will be logically organized as follows: The primary key (the Index) is the relevant law. The applied action is the secondary key. These two keys are being used for the search of relevant law + action. The rest of the record is the verdict. In our sample, the verdict for "destroying a civil house with weapons inside" (case 1), could be "the army should

compensate the house owner in the amount of 0 (zero), according to rule section N1.N2.N3", while the verdict for "destroying a civil house to make a path for the army" (case 2) could be "the army should compensate the house owner in the amount of XXX, according to rule section N1.N2.N4".

Keeping the applied rule sections along with the verdict is most important for aggregating and processing multiple verdicts. The law structure, and not the actions, is the key for the accumulation of relevant verdicts, since the actions by themselves are context-less. Using rule sections as a key for verdict management will enable cross-time correlation of verdicts, while the same rule violation considered differently in past times. Such a rule-indexed database may provide a unified evaluation system across different countries and geographies, by enabling the translation of verdicts through universal commonalities.

Inventories of legal cases are already digitized. For example, the Old Bailey on-line system (www.oldbaileyonline.org), which contains the London's central criminal court history between the years 1674-1913. Such inventories are just the first step in making the law systems accessible for machine-ethics implementations.

Legal archives of war crimes will be a primary source for lethal actions of machines that are designed for military needs. These verdicts should embed the international law and code-of-conduct regarding war actions, such as the Geneva treaty.

References

1. M. Anderson and S. L. Anderson, *Machine Ethics: Creating an Ethical Intelligent Agent*. AI Magazine Volume 28 Number 4 (2007) (© AAAI).
2. M. Anderson, S. Anderson and C. Armen, *An Approach to Computing Ethics, IEEE Intelligent Systems*. July/August, pp. 56-63, 2006. Arkin, R.C., Behavior-based Robotics, MIT Press, 1998.
3. R. C. Arkin, *Moving up the Food Chain: Motivation and Emotion in Behavior-based Robots*, in Who Needs Emotions: The Brain Meets the Robot, Eds. J. Fellous and M. Arbib, Oxford University Press, 2005.
4. R. C. Arkin, *Governing Lethal Behavior in Autonomous Systems*, Taylor and Francis, 2009.
5. R. C. Arkin, *The Case for Ethical Autonomy in Unmanned Systems*, Journal of Military Ethics, Vol. 9(4), pp. 332-341, 2010.
6. R. C. Arkin, M. Fujita, T. Takagi and R. Hasegawa, *An Ethological and Emotional Basis for Human-Robot Interaction*, Robotics and Autonomous Systems, 42 (3-4), March 2003.
7. R. C. Arkin and P. Ulam, *An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions*, IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09), Daejeon, KR, Dec. 2009.
8. R. C. Arkin, A. Wagner and B. Duncan, *Responsibility and Lethality for Unmanned Systems: Ethical Pre-mission Responsibility Advisement*, Proc. 2009 IEEE Workshop on Roboethics, Kobe JP, May 2009.
9. D. K. Ashley, *Case-Based Reasoning and its Implications for Legal Expert Systems*, Artificial Intelligence and Law 1:113-208, 1992.
10. S. A. Bringsjord, *21st-Century Ethical Hierarchy for Robots and Persons: EH*, in A World with Robots: International Conference on Robot Ethics: ICRE 2015, volume 84, page 47. Springer, 2017.
11. N. S. Govindarajulu and S. Bringsjord, *On Automating the Doctrine of Double Effect*, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).
12. D. MacKenzie, R. C. Arkin and J. Cameron, *Multiagent Mission Specification and Execution*, Autonomous Robots, Vol. 4, No. 1, pp. 29-57, Jan. 1997.
13. A. Ophir, *The Order of Evils: Toward an Ontology of Morals*, MIT Press, 2005.

14. R. Rorty, *Contingency, Irony, and Solidarity*, Cambridge University Press, 1989.
15. M. Scheutz and B. F. Malle, *Moral Robots*, in K. Rommelfanger and S. Johnson (eds.), Routledge Handbook of Neuroethics, 2018. New York, NY: Routledge/Taylor and Francis.
16. C. Strong, 1988, *Justification in Ethics*, in Baruch A. Brody, editor, Moral Theory and Moral Judgments in Medical Ethics, 193-211. Dordrecht: Kluwer Academic Publishers.
17. M. Walzer, *Just and Unjust Wars*, 4th Ed., Basic Books, 1977.

ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.016

ROBOT COMPANIONS FOR OLDER PEOPLE – ETHICAL CONCERNS

JIM TORRESEN, TRENTON SCHULZ, ZIA UDDIN, WERIA KHAKSAR
Robotics and Intelligent Systems group, Department of Informatics, University of Oslo
Oslo, Norway

EDSON PRESTES
ϕ-Robotics Research Group, Informatics Institute, Federal University of Rio Grande do Sul (UFRGS),
Porto Alegre, Brazil

The proportion of older people in the population is increasing and correspondingly, the need for support in elderly care services. As the relative number of human resources would be diminishing, technological solutions and services need to be explored. However, many ethical concerns arise when exploring these solutions and services, which we have seen in an ongoing research project – multimodal elderly care systems (MECS) – for introducing robot companions for older people. The ethical concerns focused on in this paper include privacy, security, safety and the potential lack of contact with other humans. This paper will present these issues and discuss possible ways of addressing these concerns.

1. Introduction

Would we like to be surrounded by robots rather than humans? Most would answer no to this question. However, if the question is whether we would like – with some help from robots – to be *independent* with regards to our *key needs* like personal care, eating and transportation, the answer is not as obvious (although it will vary with cognitive degradation or mobility limitations). In contrast to most enjoying to help others, the feeling of being a burden to others can be unpleasant, and we derive a sense of dignity from handling our key needs by ourselves. Thus, if a machine can help us, we prefer it in some contexts. We see this today with the Internet where we rather than asking others about how to solve a problem, seek advice on the Internet. We probably achieve things with machines which we otherwise would not get done. Thus, in the same way as Google is helping us today with information needs, robots will in the future help us with our physical needs. Of course, we still need human contact and social interaction. Thus, it is important that technology can support our social needs rather than making us more isolated. Autonomous cars may be one such measure, by enabling the elderly to go out and about more independently. Thus, such cars would support a more active social life than today where a human operated car would have to be called for if public transportation is not an option.

The multimodal elderly care systems (MECS) project* aims to create and evaluate multimodal mobile human supportive systems that can sense, learn and predict future abnormal events of elderly. A part of this will be to demonstrate the benefits regarding both performance and privacy being improved by applying sensors like cameras on a robot companion rather than having them permanently mounted in a home. These would be used for detecting falls and other non-normal situations. Using new sensor technology, we would also like to explore if it is possible to remotely monitor medical states like pulse or breathing. Rather than having elderly themselves activating their personal security alarm in the case of an emergency situation, a target

*http://www.mn.uio.no/ifi/english/research/projects/mecs/

of this project is to demonstrate *automatic* activation. Many systems for elderly have been designed but few have been adopted on a large scale. We think a key reason for this – in addition to technical limitations – is limited user involvement and few iterations of user testing. Therefore, we focus specifically on developing our systems with a large degree of user participation. In this paper, the current findings with regards to how sensing, control and user participation can impact ethical issues will be presented. That includes how to address the ethical challenges of a robot in the home. E.g. sensors in the home can record lots of sensitive information that needs to be protected so that the elderly living at home can keep their dignity and not worry about the data being misused.

The remainder of this paper is organized as follows; the next section introduces a selection of earlier works on robots and elderly. In section 3, our proposals and findings we have made with regard to various ethical concerns will be outlined, including a discussion of relevant ways of addressing the concerns. Finally, conclusions are included in section 4.

2. Background

There are a number of larger funding schemes to support the development of technology for an active and assisted living [1,2. Thus, there have been several related projects about robot companions for elderly people. This includes the CompanionAble project (2008-2012) which applied an assistive companion robot called Hector to provide care support facilities including diary management, reminder services – for example, reminders for taking medicines on time – and perform fall detection. The ACCOMPANY project (2011-2014) used a robotic companion providing services to elderly users in a motivating and socially acceptable manner. The elderly was using a tablet to interact with the robot. CORBYS (Cognitive Control Framework for Robotic Systems, 2011-2015) has a goal of making a demonstrator which match the requirements of the user at different stages of rehabilitation in a wide range of gait disorders. ExCITE (Enabling SoCial Interaction Through Embodiment, 2010-2013) had a target to evaluate user requirements for robotic telepresence employing the *Giraff* robotic platform. GiraffPlus (2012-2014) focused on monitoring activities in the home using a network of sensors, both around in the home and on the body. The robot platform was the same as in the ExCITE project. This platform is also used in the VictoryaHome project (2013-2016) targeting a support system that monitors health and safety, and facilitates social contact. Work on addressing potential ethical issues with assistive robots has also been undertaken [5. 6. . Further, the term *roboethics* has been introduced to address ethical issues related to the development and use of robots [15. . There has in recent years been taken a number of initiatives to propose possible regulations for robots and AI in the real world [12. . There is a range of ethical concerns relating to robot care for the elderly with regards to human rights and to shared human values [11. , e.g. potential lack of human contact, loss of privacy and control to name a few. Below follows a presentation of various ethical considerations and countermeasures for robot companions for elderly that we have selected and addressed in the MECS project introduced in section 1.

3. Addressing Ethical Concerns with Robots and Elderly

Focus on user needs and preferences including how a person perceives a robot are essential when developing and applying robots for elderly. Thus, both size and shape, as well as how it moves around regarding motion pattern, speed and more are important [10.]. As a part of design with user participation, it is also important working with the elderly both with talking and observation to detect issues they may have where a robot can provide a solution.

In addition, we have in our work found the following methodological approach useful for the technical development: (i) survey current available sensor technologies relevant for robots (ii) obtain sensors relevant for studies and collect data from different environments and degrees of user interaction (iii) apply current state of the art methods in feature extraction (e.g. independent component analysis, local directional pattern and deep belief networks) and classification (e.g. convolutional neural networks, recurrent neural networks and deep reinforcement learning) and combining them into novel hybrid systems for robot sensing and control [14. . (iiii) Test and verify the different robot configurations starting in lab environment and gradually moving into real user environments like elderly homes.

Below follows a description of a set of different ethical issues that have appeared in our project including proposals on how to address them.

3.1. *Privacy*

It is important to balance the privacy of the elderly against the needs for data collection for having an efficiently functioning elderly care systems [3. . Privacy in this setting regards the protection of sensitive data to avoid unwanted distribution and misuse of such data. The choice of robot sensor technology and the way sensor data in processed and potentially stored can have a major impact on privacy and vulnerability for possible misuse of data. Thus, considering and comparing these, which we have undertaken in the MECS project, are important to make progress in feasible sensor technology and processing that would be relevant out of privacy concerns. RGB camera and microphone are the kinds of sensors revealing most privacy related information. Some sensor technologies are collecting less privacy related information but may on the other hand, result in the robot not having the most accurate and effective behaviour possible but can still be relevant [12. . The lack of performance from one sensor can to some extent be compensated with using multiple different complementary sensors in combination [4. . Such a hybrid multi-sensor system can also *adaptively adjust* which sensors a robot is using depending on the given context. Relevant sensors include e.g. depth camera, force and proximity sensors, ultra-wide band radar and ultrasound sensor. Depending on the current needs for a given setting, that may be more or less privacy revealing. This can be combined with some signalling or actuation by the robot indicating to its user what and how detailed sensing that is currently undertaken. E.g. one may think of an eyelid hiding a sensor when not in use. User studies – including design with *user participation*, would here be important to assess how the user perceives a robot companion with sensors equipped. The sensor technology will, however, potentially also impact the quality of the robot-human interaction. Thus, if sensors are simpler, potentially more instructions or follow up from the user would be needed to control the robot. Similarly, by not sending sensor data over the Internet for processing in cloud resources, privacy is strengthened but limits the quality that can be provided. This is due to the more limited on-board computing power on a robot. The focus on local processing puts attention on the power consumption rather than the processing speed only. Custom hardware realization in reconfigurable hardware has shown to be preferable compared to software running on a processor [9. . That is, FPGA (Field Programmable Gate Arrays) are more energy efficient than GPUs (Graphics Processing Units) and GPUs, in turn, are more energy efficient than CPUs (Central Processing Units). However, implementing and maintaining a custom hardware design is more time consuming than regular processor software. The need for collecting and storing data is also higher when *developing* systems than when *applying* them. For development, much sensor data would be helpful to determine – by using machine learning – what sensors and features that are most effective for solving the given task and then also train the system to

provide as high performance as possible. A trained system to be applied, on the other hand, can run the sensor data processing locally and only forward some high-level status information to the caregiver. However, if there is an alarm situation, it will once again be a compromise between the benefit of being able to remotely observe the elderly and proving privacy protection. There has appeared several guidelines and regulation with respect to handling data. One is the US *Health Insurance Portability and Accountability Act* (HIPAA)[†] from 1996 targeting to properly protect health information. Another recent one is the *General Data Protection Regulation* (GDPR)[‡] designed "to harmonize data privacy laws across Europe, to protect and empower all EU citizens data privacy and to reshape the way organizations across the region approach data privacy".

In conclusion, there will always be a trade-off for a robot in a home with regards to the *conflicting objectives* of performance and privacy protection. Thus, future work should address these together and try to come up with technological solutions which provide the best possible performance while at the same time provide privacy for its user.

3.2. *Security*

The are several concerns related to security. One is related to *privacy* and possible theft and unwanted distribution of sensor data from a robot. Another is related to risk of misbehaviour of the robot in similar ways as computers can be attacked with malware. Thus, security mechanisms should be designed by analysing where to insert a protective mechanism to both handle sensor data misuse and protecting a robot from being controlled by unauthorized people. There is probably today a larger vulnerability related to privacy than robot misbehaviour but precautions for both should still be taken. Beyond regular security measures with passwords and authentication, one may also add schemes within a robot sensing and control system assessing the given context when external requests occur. That is, a robot would often have a self-aware system that is continuously updated and adapted [8. that as a part of the motion planning and communication with the outside world would consider and take into account the potential security risks, as well as other ethical challenges [12. . That is, as a part of the reasoning engine, we would add a user assessment module that can consider the current context when a remote login is received and control or data access is requested.

3.3. *Safety*

The expected upcoming wide employment of robots in our society would result in robots getting physically much closer to humans than what we are used to from protected manufacturing settings. Thus, it would be of major importance that the new robot companions operate properly for us to want them *close by*. If they hit us unintentionally or work too slowly, few would accept them. Or worse, if someone through illegal access is able to take control of a robot companion and with intention targeting to hurt us or make other damage, our trust in them would be even less.

However, there are various solutions that aim at mitigating the safety risks associated with robots. In particular, research considers mechanisms for detecting and handling safety risks as introduced in the previous section. At the same time, the control and motion planning algorithms have to be robust and well tested. Still, a robot should be able to enter a new home without extensive training and testing before it can be applied, thus, we have seen that sampling based

[†] https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html
[‡] https://www.eugdpr.org

motion planning algorithms are relevant to design and apply [7. . The research also looks into the potential trade-off between robot size, performance and safety. A small robot on the floor can be stumbled on while a bigger one may be regarded as more threatening and can also potentially represent a larger physically risk. That is, the physical design and size of a robot would impact safety issues that should be studied with user participation which is a part of our future work. Providing a robot with arm(s) and hands built from soft material would also contribute to reducing physical harm. Last, it would be important for a robot companion to contain a self-aware adaptable system that can learn about the user´s daily activities and preference and contribute in supporting these rather than introducing conflicts with them.

3.4. *Lack of Human Interaction*

As pointed out in the beginning of the paper, robots being introduced for taking care of people is a sensitive topic and often leading to many opposing the idea. Nobody likes the idea of especially elderly being left only with robots and no humans around to interact with them. However, robots can also have the opposite effect that caregivers can make robots take care of the manual work in a home to free time to talk and interact rather than doing practical work.

It is not only the robot engineers who determine how the future with robots is going to be. It will also be up to the politicians and society to decide, including on the *staffing* within elderly care when less physical work with elderly is needed. At the same time, if future robots take many of our current jobs, people in a family may, in general, have more free time, including *time* to spend together with elderly family members. And finally, today´s elderly should not worry about future robot technology. It is rather *us* who are younger, including those of us currently working with *developing* elderly care robots, that would be *confronted with* these close by robots when *we get old* in the future. Therefore, it's our own interest to make *user-friendly* robots.

4. Conclusion

The paper has introduced various ethical concerns that appear when robots are considered applied in the home care of older people. The main concerns relate to privacy, security and safety, as well as the potential lack of contact with other humans. However, there are some ways of mitigating the challenges but that may to some extent also limit the possible performance of a robot. Thus, it is important in future work to consider the different conflicting objectives being present.

Acknowledgments

This work is partially supported by the Research Council of Norway (RCN) and the Norwegian Centre for International Cooperation in Education (SIU) as a part of the Collaboration on Intelligent Machines (COINMAC) project, under grant agreement 261645 and Research Council of Norway as a part of the Multimodal Elderly Care Systems (MECS) project, under grant agreement 247697.

References

1. Active and Assisted Living Programme (2015). https://ec.europa.eu/digital-agenda/en/active-and-assisted-living-joint-programme-aal-jp
2. Ambient Assisted Living Programme (2015). http://www.aal-europe.eu/

3. A. Costa, F. Andrade, and P. Novais, Privacy and Data Protection towards Elderly Healthcare, Handbook of Research on ICTs for Human-Centered Healthcare and Social Care Services, pp. 330 – 346, 2013.
4. A. Danielsen, J. Torresen, "Recognizing Bedside Events using Thermal and Ultrasonic Readings", *Sensors*, Vol 17 (6), 1342, 2017
5. A. Ferreira M.I., J.S. Sequeira (2017) Robots in Ageing Societies. In: Aldinhas Ferreira M., Silva Sequeira J., Tokhi M., E. Kadar E., Virk G. (eds) A World with Robots. Intelligent Systems, Control and Automation: Science and Engineering, vol 84, pp 217 – 223, Springer.
6. R. Gelin (2017) The Domestic Robot: Ethical and Technical Concerns. In: Aldinhas Ferreira M., Silva Sequeira J., Tokhi M., E. Kadar E., Virk G. (eds) A World with Robots. Intelligent Systems, Control and Automation: Science and Engineering, vol 84. Springer
7. W. Khaksar, … and J. Torresen (2017). Sampling-based online motion planning for mobile robots: utilization of Tabu search and adaptive neuro-fuzzy inference system. *Neural computing & applications* (2017).
8. P. Lewis, A. Chandra, F. Faniyi, K. Glette, T. Chen, R. Bahsoon, J. Torresen and X. Yao, Architectural Aspects of Self-Aware and Self-Expressive Systems: From Psychology to Engineering, *Computer*, 48(7) (2015) IEEE Press, pp 62 – 70.
9. S. Mittal and J.S. Vetter. (2014). A Survey of Methods for Analyzing and Improving GPU Energy Efficiency. *ACM Comput. Surv*. 47, 2, Article 19 (August 2014).
10. T.W. Schulz; J. Herstad and J. Torresen (2018). Moving with Style: Classifying Human and Robot Movement at Home, In Christian Wögerer; Birgit Gersbeck-Schierholz & Steffen Gerhard Schulz (ed.), *ACHI 2018, The Eleventh International Conference on Advances in Computer-Human Interactions*. International Academy, Research and Industry Association (IARIA). ISBN 978-1-61208-616-3. 30. pp 188 – 193
11. A. Sharkey and N. Sharkey, Granny and the robots: Ethical issues in robot care for the elderly. In Ethics Inf Technol (2012) 14: 27.
12. J. Torresen, A Review of Future and Ethical Perspectives of Robotics and AI. (2018) *Frontiers in Robotics and AI*, vol. 4.
13. M. Uddin; W. Khaksar and J. Torresen, Human activity recognition using robust spatio-temporal features and convolutional neural network, In 2017 *IEEE Int. Conf. on Multisensor Fusion and Integr. for Intelligent Systems*.
14. M. Uddin; W. Khaksar and J. Torresen, Facial Expression Recognition Using Salient Features and Convolutional Neural Network. (2017) *IEEE Access*.
15. G. Veruggio and F. Operto (2008). Roboethics: Social and Ethical Implications. In Springer Handbook of Robotics (Siciliano, Bruno and Khatib, Oussama. eds). pp 1499 – 1524, Springer Berlin Heidelberg.

ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.017

APPROPRIATENESS AND FEASIBILITY OF LEGAL PERSONHOOD FOR AI SYSTEMS

BENDERT ZEVENBERGEN*

Center for Information Technology Policy, Princeton University
310 Sherrerd Hall, Princeton, NJ 08544, U.S.A.
benzevenbergen@princeton.edu

MARK A. FINLAYSON

School of Computing and Information Sciences, Florida International University
11200 S.W. 8th Street, CASE Building, Room 362,
Miami, FL 33199, U.S.A.
markaf@fiu.edu

MASON KORTZ

Harvard Cyberlaw Clinic; Berkman Klein Center for Internet & Society, Harvard University
Wasserstein Hall Suite 5018, Cambridge, MA 02138, U.S.A.
mkortz@cyber.harvard.edu

JANA SCHAICH BORG

Center for Cognitive Neuroscience & Kenan Institute for Ethics, Duke University
140 Science Dr., Durham, NC 27708, U.S.A.
janaschaichborg@gmail.com

TJAŠA ZAPUŠEK

Faculty of Law, University of Copenhagen
Njalsgade 76, 2300 København S, Denmark
tjasa.zapusek@gmail.com

The European Parliament has adopted a proposal to explore the impact of a legal personhood category for AIs, comparable to corporate personhood ("AI Personhood"). We propose that it is premature to introduce AI Personhood, primarily because (i) the scope of AI is still ill-defined, (ii) the potential economic efficiencies and distribution of gains is uncertain, (iii) the ability of existing legal structures to achieve similar ends have not been sufficiently analyzed, (iv) the moral requirements for personhood have not yet been met, and (v) it is not yet possible to assess the social concerns arising from AIs that are indistinguishable from humans. To support our conclusion, we discuss (1) the relevance of legal personhood, (2) the definitional difficulties surrounding AI, (3) currently applicable legal principles, (4) the potential benefits and drawbacks of AI Personhood, and (5) the conditions that might justify such a category in the future. We propose five specific necessary conditions—technological, economic, legal, moral, and social—for AI Personhood, but observe that the conditions have not yet been met and seem unlikely to be met soon.

Keywords: Electronic Personhood; Legal Frameworks; Liability

In February 2017, the European Parliament adopted a non-legislative resolution noting that the increasing autonomy of robots or artificial intelligence (AI) systems raises serious questions as to whether the ordinary approaches to liability are sufficient to ensure just outcomes. The resolution called on the European Commission to explore the liability implications of robots and AIs and, in particular, they raised the possibility of granting AIs status as legal persons ("AI Personhood") [1, § 59f].

*Bendert Zevenbergen is the lead author; the other authors are listed alphabetically.

Granting such "status of electronic persons"—comparable to the legal personhood assigned to corporations—is not a new idea,[2,3] and is a very real legislative possibility. This, combined with the rapid advances in robotics and AI, suggests that it is timely to carefully reconsider the arguments for and against creating a new legal category of AI Personhood.

Our conclusion is that while there may be future conditions that justify or even necessitate AI Personhood, it appears premature and probably inappropriate to introduce AI Personhood now, primarily because (i) the scope of AI is unclear, as a concept or as an artifact, (ii) it almost completely opaque what economic efficiencies will be gained, and what the distribution of economic benefits will be, (iii) we have not demonstrated that existing legal structures cannot achieve similar ends, and (iv) AIs do not yet meet the moral requirements for personhood, and are unlikely to meet them soon.

To support this conclusion, we first discuss the relevance of legal and moral personhood with respect to legal systems generally[a]. We highlight the definitional difficulties of AIs, and the problems these pose for AI Personhood. We next outline existing legal options for addressing the supposedly new problems introduced by AI, and then discuss the advantages and disadvantages of AI Personhood. Finally, we outline conditions that might justify or even necessitate AI Personhood, and conclude that these conditions have not yet been met and are unlikely to be met soon.

1. Relevance of Legal (and Moral) Personhood

Legal personhood is a construct that can be attributed at the will of the legislature, and not necessarily be driven explicitly by moral considerations. On the other hand, the commonsense concept of personhood is tied to being a human. In most legal systems there is a distinction between natural and legal persons: "Natural persons" includes all and only humans, whereas "legal persons" can exclude some humans but include non-human entities that have been deemed as needing special status. For example, societies may count corporations, nations, or political organizations as legal persons (see, e.g., Article 47 of the Treaty on European Union, or[4,5]), whereas some people—such as slaves—were historically denied legal personhood.

A legal person is an entity that can bear rights and duties,[6,7] such as the ability to own property, conclude contracts, or be sued. This is a "legal fiction" that allows non-human entities to be treated like natural persons for some aspects of the law. Without the concept of legal personhood, persons injured or harmed by a faulty product would need to find the exact person responsible within the company producing the product. Legal personhood allows the injured party instead to hold the company responsible. This is often justified by appeal to legal and economic efficiency.

Legal personhood may sometimes be founded on arguments about moral status. This usually means that an entity can suffer, or be able to reason about its own existence and moral responsibilities. Jeremy Bentham and Peter Singer have both argued that moral status is derived partially from our—or an animal's—ability to suffer.[8,9] Suffering can be physical, emotional, or financial. Kant and Regan argued that moral status is derived from intrinsic worth, our sophisticated cognitive capacity to reason about ourselves as "subjects of life".[10,11] Both definitions imply some form on consciousness. beyond the scope of this paper) By way of example, corporations can suffer financially and otherwise, though not exactly like natural persons. Further, a corporation is comprised of people who collectively can reason about the corporation's existence and its moral duties.

[a] A comprehensive review of legal personhood and specific legislation relevant to AIs is beyond our scope here. Therefore, we take a generalized and high-level view not tied to a particular jurisdiction.

2. Definitional Difficulties with AI

The first challenge when evaluating whether AIs should be assigned an existing form of legal personhood, or whether AI Personhood should be created at all, is to define what counts as "AI". The concept of AI itself is notoriously elusive, with different groups strongly disagreeing over precisely what it requires.[12] When John McCarthy coined the term "artificial intelligence" in 1955, he defined it as "the science and engineering of making intelligent machines".[13] Most definitions follow this lead by describing AI as "intelligence exhibited by machines." Common variants add that AI must demonstrate "human" or "human-like" intelligence.[14] Such definitions assume that *intelligence* is clearly defined itself, though it too is ambiguous.

We can further distinguish between narrow and general AI (a.k.a., *artificial general intelligence*, or AGI, for the latter). Narrow AI addresses specific applications, where machines often outperform humans in speed, accuracy, and efficiency. Narrow AI is already widely used. General AI, by contrast, requires intelligent behavior that is (at least) as broad, adaptive, and advanced as a human across a full range of cognitive tasks.[15] It is debatable whether consciousness is a prerequisite for intelligence, or vice versa, and also if and when general AI will be achieved. While we recognize that there is no settled definition of AI, for the purpose of evaluating AI Personhood, we define AI as *human-created digital information technologies and associated hardware that displays intelligent behavior that comes not purely from the programmer, but also through some other means.* We explicitly mention that AIs are man-made, because the designers selected the parameters within which an AI operates and learns. This also acknowledges that AIs include not only software, but also physical hardware.[14] Finally, by mentioning *some other means*, we acknowledge that AIs can develop a type of agency due to the influence of external factors on its behavior.

This definition, then, leads us to the first major problem for AI Personhood, namely: what entity specifically should be accorded the legal status? The hardware and software for an AI system can, in principle, be widely distributed, either physically or computationally.[16] The European Parliament appears to consider AIs as easily identifiable artifacts, even though this is misleading.[17] What about AIs which do not have a precisely defined embodiment, beyond the specific computer on which they temporarily reside? There is no clear answer to this problem at this time.

3. Existing & Current Law

Existing legal approaches treat AIs simply as tools. The European Parliament expressed two primary concerns with this view: that it will be difficult to establish a causal link between the harmful action of the AI and a legal person that can be sued, and that it will be difficult to identify the correct defendant when technologies from several different sources effect an AI system's behaviour. AI Personhood could solve these problems, as the European Parliament suggests, but so could other, less sweeping legal doctrines. We consider a few of those doctrines here.

A party injured by a product with an embedded AI system could hold the producer of the system liable under the doctrine of strict product liability. Under strict product liability, the supplier of a product is liable for harms caused by defects in that product, regardless of whether this was the result of negligence.[18,19] Strict product liability has two features that make it an appealing model for AI liability. First, in many jurisdictions, an injured party can prove that a product is defective without precisely identifying the defect, so long as they can show that the product was less safe than a reasonable consumer would expect and that the malfunction was not due to some external factor; the burden is then on the supplier to disprove or excuse the defect.[20] A system of strict AI liability might create a similar

presumption that any AI system that falls below some pre-defined acceptable rate of error is defective. Second, multiple parties can be joint and severally liable for the same defective product. If one component of larger system is defective, both the component manufacturer and the product assembler can be held liable.[21]

Imagine a person was injured by a defective autonomous vehicle. Under a strict product liability regime, that person could sue the vehicle manufacturer without identifying whether the defect was in a physical or AI component or proving that the defect was the result of negligence. However, the manufacturer would have an incentive to determine whether the defect was in the AI—if so, it could sue the AI developer for indemnification. This system would leave the difficult task of identifying AI defects to parties more capable of doing so than the end consumer and might incentivize development of more transparent AIs. It could also encourage AI developers and product manufacturers to apportion liability preemptively, avoiding the costs associated with *post hoc* litigation.

Product liability is generally limited to physical injuries caused by physical products.[22] How could the legal system handle, for example, a pure software financial AI that loses its clients' money? One possibility is to focus on the decision to deploy the system in the first place. The law could declare that some uses of AI are *ultrahazardous*, meaning they pose a significant risk of harm even when performed with care. Similarly, AIs could be treated as animals. Although animals are autonomous, their owners are legally responsible for them.[23] Under either approach, the law would permit recovery for at least some losses caused by AIs without requiring the injured party to prove that any particular human acted negligently or wrongfully.

Finally, we could consider *vicarious liability* for AIs. Vicarious liability is a doctrine under which one party takes legal responsibility for the conduct of another.[24] If an AI system committed a tortious act, liability would be determined as if the owner had committed that act. The owner could not evade responsibility by claiming lack of knowledge or intent.

4. Advantages of AI Personhood

Even though we have argued that there are ample existing legal mechanisms that could potentially address the issues that AI Personhood is intended to solve, the creation of a new AI Personhood category is nonetheless a real possibility. We see two main potential benefits. First, there are potential instrumental advantages to AI Personhood, which are suggested by analogy to corporate personhood. Corporate personhood allows a connected group of persons—though potentially distributed in time or space—to pool resources and centralize risks. This pooling of resources can be necessary to spur large-scale innovations or to take advantage of economies of scale. In turn, the economy and society in general derives a benefit. From a legal point of view, corporate personhood allows single organizations to be held liable for harms without the need to identify a responsible individual. Legal efficiency is achieved because it allows plaintiffs to sue the organization directly without going through a lengthy, expensive, and arduous process of identifying the specific individuals responsible. Economic efficiency is achieved by the pooling of resources to increase productivity, while creating legal certainty improves the efficiency of operation.

A second benefit is based on morality coupled with a potential technological trajectory: If, in the future, a general AI system is developed that is indistinguishable from a person, by what argument do we deny that system the same rights as a human? More strongly, if an AI can be shown to have real consciousness, suffer real pain, or be truly independent, a majority of the population might feel morally compelled to grant the AI the same rights and responsibilities as humans. While it is clear that AIs are not yet at this level (and it is unclear if or when they will reach it), it would be unwise to dismiss this possibility completely. At

that point, the question may become less an issue of legal or economic efficiencies as it is of "human" rights.

5. Disadvantages of AI Personhood

There are many possible disadvantages of AI Personhood. We list four scenarios that we see as most likely in the near term. First, the European Parliament suggest creating a collective insurance fund to cover damages arising from AIs. However, the technological trajectory of AI is uncertain and unpredictable, and it is therefore unwise to construct financial compensation resources today to meet as yet unknown future needs.

Second, AI Personhood would allow producers and owners of AIs to shift liability to the artifact itself. This will disincentivize investment in adequate testing before deployment. AI Personhood could thus result in an unsafe environment wherever AIs are deployed.[25]

Third, it will be difficult to bring proceedings against AIs or hold them to account. A corporation may employ lawyers or seek outside counsel. AIs do not (yet) have the capacity to argue their case in court, appoint a lawyer to represent their interests, or engage meaningfully with a plaintiff to reach a settlement; furthermore, these capacities do not seem likely soon.

Finally, AIs do not yet have the capacity to suffer, and it is unclear if it is possible to program or develop empathy digitally, such that an AI would meaningfully understand suffering in others. Further, an AI system cannot today interpret it ethical responsibilities on a contextual basis, nor is it intrinsically aware of its own existence.

6. Conditions for AI Personhood

From the above, we distill four conditions for AI Personhood.

Technological We need to be able to delimit the boundaries of a particular AI system, as AIs can integrate and depend on many external systems for their functioning.

Economic If AI Personhood allows for increased innovation and economic growth, we must identify the beneficiaries and how the gains benefit society. Negative externalities and consequences should be understood and accounted for. Instrumental economic reasons must be scrutinized from a diverse range of perspectives.

Legal AI Personhood would be a far-reaching change in society that must not be taken lightly. Arguments from legal efficiency would require evidence that the current law is insufficient. Similarly, a claim that current law retards the development of beneficial AIs must be carefully assessed. Significant justification should be required to enact such a fundamental change to the legal system, and great care should be taken that AI Personhood is not abused by powerful interests.

Moral AIs must begin to function like current legal persons (i.e., individuals, corporations, and nations). In line with Bentham and Singer, we agree that the ability to suffer in some form is essential. On the other hand, we would also argue that it may be considered immoral to create an AI system that can suffer in the first place. In addition, we agree with Kant and Regan that some form of intrinsic worth, such as the ability to reason about one's own existence and moral duties, is critical.

7. Conclusion

In our view, none of these four conditions are met today. The technological trajectory also does not point in the direction that these conditions will be met soon. We can imagine that in the far future AIs may become much more like people, to the point where we are morally compelled to grant them rights and responsibilities. In fact, several movies and

science fiction stories have allowed us to imagine this technological trajectory. However, this trajectory is more difficult to foresee based on the current state of AI research. We therefore do not think that a speculative possibility should affect or legislative decisions and resources today.

8. Acknowledgements

The main ideas of this paper were discussed at a workshop on AI Personhood at the Princeton Center for Information Technology Policy (CITP) on May 11 & 12, 2017. The workshop attendees included the authors, Peter Asaro, Joanna Bryson, Thomas Burri, Vincent Conitzer, Ed Felten, Brett Frischmann, John Havens, Joanny Huey, Konstantinos Karachalios, Ugo Pagallo, Joel Reidenberg, and Yan Shvartzshnaider. Ugo Pagallo contributed significantly to prior drafts of this paper. Mr. Zevenbergen was partially supported by a CITP fellowship, and Dr. Finlayson by ONR contract No. N00014-17-1-2983.

References

1. European Parliament, Resolution (2015/2103(INL)) of 16 Feb 2017.
2. L. B. Solum, *N.C. L. Rev.* **70**, 1231 (1992).
3. S. Chopra and L. White, Artificial agents: Personhood in law and philosophy, in *Proc. 16th Euro. Conf. on Artif. Intell.*, (Valencia, Spain, 2004).
4. J. R. Crawford, *The Creation of States in International Law* (Oxford University Press, Oxford, 2006).
5. S. K. Ripken, *Fordham J. of Corp. & Finan. L.* **15**, 97 (2009).
6. J. C. Gray, *The Nature and Sources of the Law* (Columbia, New York, 1921).
7. J. W. Salmond, *The Theory of the Law* (Steven & Haynes, London, 1902).
8. J. Bentham, *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation* (Oxford, Oxford, 1789).
9. P. Singer, A utilitarian defense of animal liberation, in *Environmental Ethics: Readings in Theory and Application*, eds. L. P. Pojman and P. Pojman (Cengage Learning, Boston, MA, 1998) pp. 39–46.
10. I. Kant, *The Metaphysics of Morals* (Cambridge, Cambridge, UK, 1996).
11. T. Regan, The case for animal rights, in *Advances in Animal Welfare Science 1986/87*, eds. M. Fox and L. Mickley (Springer, Amsterdam, 1987) p. 179.
12. S. Legg and M. Hutter, A collection of definitions of intelligence, in *Proc. 2007 Conf. on Adv. in Artif. Gen. Intell.*, (IOS, Amsterdam, 2007).
13. J. McCarthy, What is artificial intelligence? (2007), `http://www-formal.stanford.edu/jmc/whatisai.pdf`. Last access Nov. 15, 2017.
14. M. Scherer, *Harv. J. L. & Tech.* **29**, 353 (2016).
15. B. Goertzel and C. Pennachin (eds.), *Artificial General Intelligence* (Springer, Berlin, 2007).
16. T. Hwang, Computational power and the social impact of artificial intelligence (2018), doi:10.2139/ssrn.3147971. Last access 15 Jul 2018.
17. C. E. Karnow, *Berk. Tech. L. J.* **11**, 147 (1996).
18. Council of the European Union, Dir. 85/374/EEC, vol. L 210, 07/08/1985.
19. American Law Institute, *Restatement (3rd) of Torts: Products Liability.* (American Law Institute Publishers, Philadelphia, 1998).
20. L. Sterrett, *Mich. St. Int. L. Rev.* **23**, 885 (2014).
21. M. S. Shapo, *Corn. Int. L. J.* **26**, 279 (1993).
22. V. R. Johnson, *Washington and Lee Law Review* **66**, 523 (2009).
23. American Law Institute, *Restatement (2nd) Torts.* (American Law Institute Publishers, Philadelphia, 1977).
24. American Law Institute, *Restatement (3rd) of Agency.* (American Law Institute Publishers, Philadelphia, 2006).
25. J. J. Bryson, M. E. Diamantis and T. D. Grant, *Artif. Intell. & L.* **25**, 273 (2007).

ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.009

FRAMING RISK, THE NEW PHENOMENON OF DATA SURVEILLANCE AND DATA MONETISATION; FROM AN 'ALWAYS-ON' CULTURE TO 'ALWAYS-ON' ARTIFICIAL INTELLIGENCE ASSISTANTS

MARTIN CUNNEEN and MARTIN MULLINS

Abstract. Online connectivity now defines our 'information civilisation' and presents many benefits and risks. The dynamics of these multi-layered risk/benefit relationships are complex, but what is common throughout are risks relating to metrics of increasing values, from number of users connected, types of connectivity, time users spend connected, the number of connected devices, and the increase in user data harvesting. The online phenomenon presents an increasingly complex risk phenomenon. Fortunately, research confronts many of these risk contexts, so much so there are many growing narratives of both benefits and risks regarding online connectivity. The article focuses on one particular narrative concerning the risks of the connected online phenomenon. For the ease of discussion, we use Sherry Turkle's 2006 work the "Tethered Self" as the start of the online connectivity and risk narrative. Turkle framed some of the risks of increasing connectivity, under the title of the "always-on culture". The narrative has grown in recent times with the addition of the internet of things as another medium of connectivity, consisting of numerous forms of "always-on devices". The article maintains that the growing popularity and development of artificial intelligence assistants presents another evolutionary sequence of the always-on narrative. Furthermore, the narrative now moves from user-controlled connectivity, third party connectivity to connectivity mediated through artificial intelligence assistants/agents. The article aims to interrogate and contribute to the risk framing of artificial intelligence assistants by situating the technology in the always-on narrative.

1. Introduction

Developments in artificial intelligence (AI) will pose substantial risk governance issues for society in the coming years. We are already seeing a move towards ubiquitous network connectivity with its associated issues which may be described in terms of, to use Turkle's terminology, "the tethered self." [50]. Here we see the self as a user, is somehow compromised as it struggles to deal with a more or less constant interface with the digital world. The divide between offline real world and online digital world is no longer clear. That said, the current situation does allow for some elements of agency in that consumers can opt in and out of the digital world, but this is becoming increasingly unclear, uncertain and we claim unattainable. We maintain that over time we will move from a situation where citizens retain the ability to navigate in and out of the digital world, to one where the dominance of the IOT will tend to compromise that freedom and then finally to a situation where the self is not only tethered to the digital world but through AI assistants (AIAs) is effectively guided through it. The implications in terms of personhood, human agency and power asymmetries are enormous. The process of monetarisation of data is already well under way and has created strong momentum for this move through phrases 1-3 to take place. At the same time this process poses unique challenges in terms of risk governance. What is at stake are current paradigms around informed consent and human agency. The paper claims that in framing the risks associated with these issues concerning the three phases of - "always-on", the ubiquity of IOT and the presence of AI assistants, we will bring greater clarity to the many actors faced with creating systems of risk governance. The contextualisation of the three phases of evolving connectivity draws attention to an increasingly complex connectivity landscape and governance regime. The contextualisation is beneficial in elucidating the changing landscape of relations and risks between the diverse array of actors.

AIAs are a state-of-the-art example of online connectivity, sustained, mediated and filtered through a prism of cloud-based AI. The paper attempts to conceptually frame AIAs in the context of three potential risk metrics; (1) risk regarding Sherry Turkle risk framing as the "always-on culture" [53, 50, 54, 52] and Catherine Middleton's conception [35] and (2) risk framed in terms of the internet of things and "always-on" technologies/devices [17]. Both are inherently related conceptually but each present different technological relations between the user, how they connect, what data is harvested and what data the user is aware of generating. Therefore, the context of always-on in each case present important differences in meaning. We defend the framing of both, as intrinsic to the new phenomenon of surveillance capitalism [62] and data monetisation [24, 23], which presents the third risk metric. With increasing popularity and sophistication, AIAs will present another medium of connectivity. What is particularly different regarding this medium concerns the primary function of harvesting user data. AIAs will gather both user and environmental data from each living/social space they are placed in. As sensory technologies improve, AIAs will have audio and video data feeds, face and voice recognition, as well as other abilities to support more targeted data gathering. Moreover, it is expected that Amazon will add more advanced video capabilities to Alexa devices by late 2018 [2]. Face recognition and behavioural analytics will undoubtedly inform future devices, but it is the currently unknown future new uses of data that also poses significant risk. Most importantly, the third risk concerns the many questions relating to how gathered/harvested data is stored, used for analytics, data wrangling, user behaviour studies and ultimately used to generate profit [23, 24]. All of which are dependent upon the service provider's inference that the data users generate on privately owned platforms, services and infrastructure is owned by the providers and not the users. The amount of data generated by the digital world doubles every two years and it is expected to consist of 40 zettabytes of data by 2020.[1] A great deal of this data will be generated by users, IOTs and AIAs. Accordingly, there are many questions to consider from who owns, controls, is accountable to who benefits from this data?

2. Risk One: The Always-On Culture

The internet offers an online world of digital domains and digital spaces to meet, access information, and services. The basic format of information sharing, and communication largely remains the same as outlined in the first website created in 1989.[2] Nearly thirty years on and the virtual online world now has one in four people using social media[3] to connect and meet others. Daily online social engagements and transactions amount to several billion, the number of users is increasing every day, it is estimated 4.5 billion Facebook users daily like a post.[4] Online connectivity has become something we are obliged to use, a social norm that is becoming compulsory. This is largely a result in the change in connectivity to mobile smartphone technologies which are now inexpensive to use and increasingly support inexpensive online or free access. One reason for this change concerns how user numbers add value to platforms; this has been the case for some time in relation to marketing, advertising, click bait, redirects and so on. Facebook for over a decade continues to offer a free to use service that is built on a model of creating data monetisation from massive amounts of online users generating behavioural data. However, a new phenomenon has evolved from the value of users that is less transparent than a targeted advert or site redirect, it concerns using devices as machines to harvest user data to profile the user and to create data insights that can be used in-house or sold to third parties [14].

[1] See: https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm
[2] See: http://info.cern.ch/hypertext/WWW/TheProject.html
[3] https://www.medicalnewstoday.com/articles/275361.php
[4] https://www.webpagefx.com/internet-real-time/

2.1. The Tethered Self

In 2006, even before the exponential growth of smartphones and mobile connectivity, many perceived the dangers and risks of an emerging always-on culture. Since that time Sherry Turkle has continued to develop this narrative of always-on connectivity as the always-on culture [53, 54, 50, 52]. So much so Turkle argued there were significant psychological risks associated with the always-on culture beginning to establish itself [50]. Since 2006, the always-on culture and the ubiquitous use of online platforms have presented numerous risks and become important governance issues [30, 7, 1, 21]. To date, the online world has largely been a domain defined by a boundary of user connection and group connectivity. As individuals, we possessed the important capacity to log out, shut down a device, and disconnect. With mobile connected devices such as laptops, smartphones and wearables, we have become what Turkle refers to as always tethered and hooked into the network by means of our desire to remain in the loop and show our availability [53]. The always-on culture in part consists of the human desire to be socially active, to be in the social loop and to remain socially informed. The connection was one sustained by this social desire. Accordingly, the ability to disconnect was a form of empowerment, a choice to say no I have had enough of Facebook, prompts and email. It was a means to return to the more natural environment of life without being hooked in. Disconnection for most did not mean social isolation, it merely meant refocusing on other non-digital social networks. However, with growing digitization, this ability, this choice to turn off and disconnect has been eroded for some time. With increasing use and ease of access, we have become psychologically hooked into online forms of socialization, and this is what Turkle was criticizing in much of her work.

2.2. Risk and Governance of Always -On and Disconnection

There are numerous governance strategies employed to combat the identified risks associated with always-on connectivity. These are reflected in the right to be forgotten (RTF) movement [43, 38] and the right to disconnect (RTD) [48] [22], regarding supporting the ability to disconnect from work and to sustain a work/life balance. It is sometimes phrased as the right to be forgotten after work hours "le droit de la de´connexion, or the *"right to disconnect*"" and some corporations have supported employees to disconnect and even delete out of hours emails [47].[5] The challenge with emerging technologies such as always-on and ubiquitous always-on devices, embodied AI and AIAs, is that the ability to disconnect will no longer be straight forward or achievable. The challenge to disconnect will be beyond the capacity of users; accordingly, it is becoming a political question, it is becoming a question of governance. This is evident when one considers the case of Papua New Guinea and its state-wide disconnection from FB.[6] The RSPH report [4] states that "91% of 16-24-year-olds use the internet for social networking", while this is a surprising figure, also surprising is the claim that follows; "Rates of anxiety and depression in young people have risen 70% in the past 25 years". The report claims that the move to virtual domains and online social networking has had a significant negative impact on young people regarding a 70% increase in depression and anxiety. The report is an important medium highlighting the dangers of the always-on social media culture that defines young people. It is also important in recommending numerous key responses to this difficult scenario. In a study carried out by Anxiety UK, the charities chief executive, Nicky Lidbetter maintains that participants identified the challenging scenario of breaking the social media use cycle and ultimately this could only be achieved by turning off the device.[7]

The always-on culture represents two distinct forms of risk arising from always-on

[5] See: https://www.irishtimes.com/opinion/editorial/reply-to-all-the-right-to-disconnect- digitally-1.2927773

[6] See: https://postcourier.com.pg/shutting-facebook-png-reality/

[7] See: https://www.medicalnewstoday.com/articles/247616.php

connectivity, the first concerns the undermining of a work/life balance and the second relates to potential psychological risks associated with the change from real-world socialization to online socialization. The former is summarized as "*identification of work intensification and work extension practices*" [35], and the latter relates to concerns forwarded by re- search such as the Royal Society Public Health (RSPH) report identifying risks of social media use and young people [4,45]. The above highlights a part of the many societal and ethical tensions that relate to online connectivity regarding both commercial and domestic communications. On the one hand, accessibility for all seems meritorious and on the other, it could pose negative outcomes for those who are supported to connect via the stress associated with the always-on culture, the anxiety of social networking and the erosion of a distinct work/life balance. The need to address such dilemmas and tensions regarding connectivity is now legitimized by the recent efforts of several states to provide governance measures to mediate between the demands of society for always-on connectivity and social networking, the demands of business to harness the benefits of always-on connectivity as well as the more recent phenomenon that has evolved from the always-on culture and social networking regarding data monetisation.

3. Risk Two: Always-On Devices

Turkle partly anticipated the more penetrating social phenomenon of always-on connectivity, further strengthened by a range of devices, from wearable tech such as fit- ness trackers to external third-party devices that proliferate our social, domestic and work spaces. These devices are best described by the category of the internet of things (IOT). Along with geotagging, our domestic spaces are increasingly becoming network spaces via a myriad of devices from fridges, pet feeders and speakers, all connected. Businesses are also seeking to adapt connected technologies to wearable technologies, in order to track employee activities. Collectively, as users, our data consists of mobile connectivity via smart phones, along with the IOT, which includes increasing amounts of context specific environmental data. The addition of the IOT to the always-on narrative, changes the paradigm of always-on from a paradigm of mobile user connectivity, to a new paradigm that incorporates external networked devices. Both point to a paradigm of connectivity that moves beyond the limits of user connectivity and the capacity to disconnect by switching off or simply disconnecting by leaving the smartphone behind. Turkle's "always-on culture" concerns the societal and psychological phenomenon of online connectivity and the risks it presents. Whereas, Gray's concept of always-on refers to many devices including "*mobile phones, televisions, cars, toys, and personal home assistants—many of which are powered and enhanced by speech recognition technology*" [17]. Gray is critical of the term "always-on" to refer to a range of devices, as there are important differences between devices relating to the extent of the data the devices have access to. But both Turkle's and Gray's description of always-on reflects a world of increasing connectivity, with the increasing ubiquity of connected devices, social platforms and network access.

3.1. IOT, Connectivity and Data Collection

Gray's focus highlights the growing network of connectivity that presents a phenomenon where it is increasingly difficult to exercise choice to switch off, disconnect or unplug [17]. The infrastructure supporting network connectivity is now centred on wireless connectivity and mobile apps uniting user identity across software platforms and devices [5]. What is now emerging, from the proliferation of wireless network connectivity, the growth of global online platforms, the massive amounts of user data available, supplemented by new data harvesting devices, and the tools to analyse the raw data into even larger data sets, comes new opportunities for data controllers [24, 16, 26]. Big data and AI via always-on devices are designed to support more efficient data monetisation models of commerce [3, 23, 10]. Actors monetise data by means of a service/user agreement to support a legal right to access user data for in-house or third-party analytics. Such agreements are an example of users supporting data monetisation because of

their own data disorientation. This is echoed by Gray:

> *"There is no doubt that the increasing prevalence of voice integration into everyday appliances enables companies to collect, store, analyze, and share increasing amounts of personal data. But what kinds of data are these devices actually collecting, when are they collecting it, and what are they doing with it?" [17]*

Smart technologies, whose operation and functionality are often more than they appear to users, present important examples of innovation that require risk/benefit analysis. This is because the risk/benefit relationship is complex, it is no longer clear what benefit the manufacturer/operator/service provider receives by supplying the product. It is no longer clear to users what the purpose of the technology is for or how the corporation gains profit from the product. For example, if the data a user generates by using a technology benefits the service provider and manufacturer by using the data for monetisation, should the user be fully informed in an explainable format how their data generates profit for the corporation? Such contexts of dual-use data technologies highlight how the front end provides convenience and domestic benefits. Whereas, the front-end service is secondary to the true functionality of the AIAs, as the primary function is not providing a service to users rather it is focused on gathering data generated by users to provide financial benefits to the service providers.

3.2. IOT, Connectivity and Risk

The always-on devices are reformatting the relationship between user and technological risk, as has been defined by the three examples. Accordingly, there is a need to understand the risk/benefit analysis of products that harvest user data, as always-on devices are now designed to. The ability to mitigate risk can prove to be valuable to industry and commerce. Risk management is now common practice and works alongside innovation and societal anticipatory re- search. Accordingly, risk provides fundamental knowledge metrics that constitute an intrinsic part to anticipatory governance research and governance systems. The utilisation of risk as a knowledge domain can prove fruitful given that a main part of its utility is the need to frame phenomenon in terms of potential harms/benefits metrication. This is particularly intuitive and informative in the context of technology, especially consumer technologies that are sold to consumers as offering benefits, with little attention given to the possible risks or harms associated with use. So much so that the question of risk is seldom stated, unless it is specified by law, as is the case with the identification of possible harms. For legal determinations to become mandatory, it is necessary that a scientific burden of proof is attained but this takes time and the pace of technology has confronted systems of governance with a pacing problem [31]. This is why in recent times, technology ethics and risk has evolved to become a key knowledge source to anticipatory research and governance.

4. Risk Three: Artificial Intelligence Assistants; Data Surveillance and Data Monetisation

Connectivity now consists of user connectivity via our always-on smart phones, third party connectivity such as facial scanners, security cameras, and a host of other networked devices via the IOT. With increasing network infrastructure, smartphones with numerous forms of connectivity in one device, third-party devices, and cross platform/device user profiles, connectivity consists of a complex multi-strand mesh phenomenon. Accordingly, we no longer make the decision to connect, we are now just connected by default. Pepita Hesselberth addresses this issue as the dilemma of connectivity and losing the capacity to disconnect [22]. The challenge of connectivity/disconnectivity is no longer user centred, it has moved beyond

the capacity of the user to achieve disconnection. The online phenomenon is embedded into society, it now consists of billions of always-on devices (IOT). This expansive social mesh of network connectivity framed as the always-on phenomenon of users and devices, is now undergoing an important development. The devices that define this always-on phenomenon are being upgraded with intelligence by adding the functionality of cloud-based AI assistants. This is particularly evident in the domestic market with the growing popularity of AI assistants such as Siri, Cortana, Alexa, Google assistant, Bixby and emerging technology such as Google's Duplex, available on many phones, wireless speakers and numerous other devices [56, 2, 29].

4.1. Framing Artificial Intelligence

AIAs present an upgrade to the network phenomenon that supports the intelligent analysis of user data, user experience and user behaviour profiling. Already there are examples of functions that present potential risks and challenges, from Google's Duplex[8] deceiving a receptionist into thinking it was a human making the booking or Amazon's Alexa transferring voice data to third parties. The challenges and risks range from regulation gaps, conceptual confusions to hardware or programming faults. This changing phenomenon of socially embedded AI technologies, present many challenges, especially regarding how to conceptually frame the technologies. There are many applications of AI and to avoid confusion between the different intelligence contexts, we claim it is beneficial to move away from a general categorization of AI and instead contextualize the specific AI technologies in existent technological narratives. This is in order to frame AI technologies in a context of meaning where the technologies will be used. By doing so the specific contexts of application can be framed and interrogated. It is now important to assess the societal, ethical and legal impacts, risks, tensions and challenges that the specific applications of AI technologies present [41].

4.2. Framing AIAs and Risk

John Danaher describes the human/machine engagement and functionality of AIAs, in the context of cognitive out sourcing that presents a complex relational framework in need of both conceptual and ethical interrogation [14]. Whereas, Sherry Turkle maintains that there are more emotionally centred issues that need investigating, this is evident in the way we engage emotionally with forms of AI, from assistants to robots. What appears like a sophisticated emotional response from an AIA may also have risks relating to how users are determined by the response. For example, if an AIA makes emotionally weighted statements relating to being turned off or not being turned out or used enough, this presents both ethical and risk contexts relating to user engagement. Accordingly, many examples of embodied AI and AIAs use will present new relationships, that we perhaps as users of technology misunderstand [52]. There are also concerns as to how AIAs could be used to support behavioural and emotional responses from users [13], develop a digital dependency in replacement to human engagement [15], and present cognitive degeneration [14]. Many of these concerns can be situated in the existent literature regarding the evolution from analogue to digital technologies.

There are three key typologies of risk relating to the emerging phenomenon of always-on AIAs that can be brought together to forward a beneficial risk narrative to interrogate AIAs. The first two relate to known risks concerning the phenomenon of what is identified as the always-on culture [53, 51] and the more recent phenomenon of always-on devices [17]. Both of these risks are primarily framed in terms of risks relating to online connectivity, time connected and the inability to dis- connect. A large corpus of literature relates to both contexts of always-on risk. So much so that, we claim that when brought together, these two examples constitute an important technological narrative that frames user and societal risk in terms of online

[8] See https://www.bizjournals.com/sanjose/news/2018/07/06/google-duplex-call-center-customer-goog.html

connectivity. The third risk metric concerns AIAs as an emerging risk [44] phenomenon that presents additional risk scenarios, from AI risk exposure to decision processing, information retrieval and bias, to filtering online experience. AIAs are beginning to present a socially embedded example of AI, that users are typically unaware of the many potential risks to using the technology, the dual use context and the volume of data the assistants will analyse and harvest [39]. Accordingly, AIAs present a host of potential risk ranging from changing the HMI of user connectivity to introducing a dual use technology that not only offers benefits to users by means of its functions but also represents an important source of revenue to the service providers by collecting and harvesting user data for both data surveillance and monetisation purposes. We argue that this situation can in part be addressed by contextualizing AIAs in the established and well documented narrative of always-on risk. This supports framing AIAs in terms of identifiable risks regarding connectivity, time connected and the capacity or right to disconnect, with new risks presented by AIAs in the form of user data surveillance and data monetisation. Collectively, the metrics of identifiable risks of the always-on culture and always-on devices, framed alongside the new risks of data surveillance and monetisation, present the new risk phenomenon in a context of a technological narrative that addresses risk. AIAs and recent innovation require that we re-examine and update the always-on narrative in terms of risk. This is supported when one considers that connectivity not only crosses numerous devices, places and networks, it is now with the advent of AIAs intelligently mediated. We have framed online risk in terms of connectivity and listed numerous metrics relating to it. The actual form of connectivity, the ease of connection, the time connected, how connectivity is intrinsic in determining online user experience, how it relates to what user data is available to service providers and third parties, and the capacity to disconnect, are all key risk metrics relating to user connection. In each one of these risk metrics, relating to the context of an always-on risk framework, AIAs present stronger risk metrics. In addition to this, AIAs also present several important additional risk metrics. One important example relates to how the form of connectivity changes from a HMI, that centres on physical user actions regarding hand motions of typing, clicking or swiping, to natural language processing technology. The change to a paradigm of voice interaction is an important example of a changing risk metric that presents many difficulties.

4.3. Understanding Risk Through a Lens of Connectivity

The form of connectivity has changed over the past three decades, it is no longer activated by our use, we no longer dial in, we automatically by default connect to Wi-Fi, or mobile networks, and we typed or swiped to engage online. In 2011 this paradigm model of connectivity began to change with Apple's introduction of Siri, an always-on voice assistant that could receive voice commands and deliver information via voice, a user connects online via a cloud-based artificial intelligence assistant. The growing phenomenon of artificial intelligence assistants needs interrogating. Although, the metric of connectivity remains key, it is now undergoing significant change. Connected society started to change in an important novel way, our once latent online connections are increasingly mediated through a prism of artificial intelligence. An important and challenging question to consider is how do we understand and conceptually frame a technology that is designed to have a dual use? How are we to understand the risks of using a technology, if the front-end use of it is simple and entertaining but the back-end use it designed to retrieve personal user data. As citizens, as users, and as members of a society, there are societal, ethical and legal frameworks to protect us from harms. Whether it is freedom of expression, privacy, the right to be forgotten, data access, ownership or the right to disconnect, governance struggles to keep pace with technological and emerging data centred innovation. Accordingly, this scenario presents an opportunistic period to data monetisation actors, that benefit from this governance vacuum and lack of informed user consent. We are as users now tethered to not just a social network, but a network wherein the tether is open to control,

mediation and bias from examples of AI technologies, that are designed to be more akin to sales assistants than personal assistants. User data is becoming an increasing lucrative commodity, and so data surveillance and data monetisation are defining the online experience of uses. The always-on culture has become a necessary component of a lucrative data monetisation corporate culture, largely operating in a data wild west[9] with a cohort of data wranglers creating financial gain/profit from user data.

5. Conclusion

It is apparent that systems of governance struggle to respond to emerging technologies, the advent of anticipatory governance is testament to the lag time that exists between advances in science and regulatory responses. Given the significant change that innovation presents to society, it is difficult to frame the innovation accurately, create informed metrics of anticipated impacts and possible risk to respond to the innovation in a timely and accurate manner [31]. This scenario is unwelcome, given that some actors see this as an opportunity to exploit the lack of legislative controls. So much so it's often described as a lawless wild west of data [26, 59], wherein opportunity and profit is in part, built upon the premise that it's open season on data as it is not illegal or unethical until it is determined by authorities to be so. With all user data stored and claimed by service providers as theirs to use (until legislation determines otherwise), we are confronted with the wild west of data. Although users agree to terms of service (TOS) or user agreements before use, it remains that this agreement is not forged on transparency, explainability and informed consent. From the standpoint of data surveillance and the monetisation model, it is important to recognize the business context; social media platforms cost money to develop, to sustain, to create a secure environment and to support functionality. The model of providing services to users, in order to acquire user data, is a business model that will remain, as long as there is an opportunity to access data. Therefore, the data users generate on platforms is a commodity that is valuable to data monetisation models of business. A key risk confronting this model is user disconnection which is the most appropriate means of users mitigating always-on connectivity.

There are evident challenges concerning; (1) how numerous different actors engage and understand the technology and, specifically, it's social agency, (2) how the technology is designed and developed in an informed, transparent manner that can provide an important metric of explainability to users, (3) how the technology is governed to ensure that risks are accurately considered, and (4) how end users understand and engage with the technology, regarding making informed decisions as to the risk, benefits and limitations of the technology. In short, embodied AI technologies, such as AIAs, present more significant challenges to how the technologies are engaged. The article has identified the need to situate AI technologies, such as AIAs, in the context of a continuum of the development of the digital society. Our purpose is to update and develop existing narratives, so as to provide more accurate three-part risk frameworks that can promote more timely and accurate governance. It can also contribute to developing a more informed risk awareness in relation to users developing a more risk informed framing of the technology. In closing, it is important to frame AIAs in the context of risk, and with cognizance that users are by the very design of the technology, confronted with a built-in bias focused on data monetisation. It is also important to see the advent of AIAs as part of continuum and as we move across that continuum from more contemporary debates around the "always-on" toward AIAs as a dominant vector of our engagement with the digital world, the risk governance challenges will become more acute.

[9] https://www.telegraph.co.uk/technology/2018/05/19/gdpr-wild-west-rush-data-law-digital-age/ https://stratechery.com/2018/techs-two-philosophies/ and https://www.theguardian.com/commentisfree/2018 no-moral-code-racist-ads-cambridge-analytica-technology-ethical-deficit

References

1. National comparisons of risks and safety on the internet, http://eprints.lse.ac.uk/39608/.
2. Amazon press release (2017), http://phx.corporate-ir.net/phoenix.zhtml?c=176060&p=irol-newsArticle&ID=2303267
3. Big Data: A Challenging Technology. International Journal of Recent Trends in Engineering and Research **3**(6), 227–233 (2017)
4. Royal Society for Public Health (RSPH) submission to inquiry on the impact of cyber- bullying on social media on children and young people's mental health (2017), https:
//www.rsph.org.uk/uploads/assets/uploaded/ec7a4710-18be-463f-a3f8f5e3fd52c367.pdf
5. Integration of IoT, Transport SDN, and Edge/Cloud Computing for Dynamic Distribution of IoT Analytics and Efficient Use of Network Resources,. J. Lightwave Technol **36**, 1420– 1428 (2018)
6. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Ayyash, M.A.M.: "Internet of Things: A Sur- vey on Enabling Technologies, Protocols, and Applications," in. IEEE Communications Surveys & Tutorials **17**(4), 2347–2376 (2015)
7. Berson, I.R., Ferron, B.M., J.M.: Emerging Risks of Violence in the Digital Age. Journal of School Violence **1**(2), 51–71 (2002)
8. Born, R.: Artificial intelligence: The case against. Routledge (2018)
9. Bringsjord, S.: Psychometric artificial intelligence. Journal of Experimental & Theoretical Artificial Intelligence **23**(3), 271–277 (2011)
10. Ceusters, W., Hsu, C.Y., Smith, B.: Clinical data wrangling using Ontological Realism and Referent Tracking. pp. 1327–27. Houston (2014)
11. Chen, B., Wan, J., Shu, L., Li, P., Mukherjee, M.Y., B.: Smart Factory of Industry 4.0: Key Technologies, Application Case, and Challenges. IEEE Access **6**, 6505–6519 (2018)
12. Chui, M., Manyika, J., Miremadi, M.: where-machines-could-replace-humans-and-where-they-cant-yet (2018), https://www.mckinsey.com/business-functions/digital-mckinsey/our- insights/where-machines-could-replace-humans-and-where-they-cant-yet
13. Coeckelbergh, M.: Moral appearances: emotions, robots, and human morality. Ethics and (2010)
14. Danaher, J.: Toward an Ethics of AI Assistants: an Initial Framework. Philosophy & Technology **:**, 10–1007 (2018)
15. E., O., D, R.: Towards a Sociological Understanding of Robots as Companions. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering **59** (2011)
16. Endel, F.P., H.: Data Wrangling: Making data useful again, IFAC-PapersOnLine, ISSN 2405- 8963
17. Gray, S.: Always-on: privacy implications of microphone- enabled devices (2016), https:
18. Gunkel, D.: Social Contract 2.0 : Terms of Service Agreements and Political Theory. Journal of Media Critiques **1**, 145–168 (2014)
19. Gunkel, D.J.: The Machine Question: Critical Perspectives on Ai, Robots, and Ethics. MIT Press (2012)
20. Gunkel, D.J.: Computational Interpersonal Communication: Communication Studies and Spoken Dialogue Systems, (2016)
21. Hasebrink, U., Goerzig, A., Haddon, L., Livingstone, K.V., S.: Patterns of risk and safety online: in-depth analyses from the EU Kids Online survey (2011), http://eprints.lse.ac.uk/ 39356/.
22. Hesselberth, P.: Discourses on disconnectivity and the right to disconnect. New Media & Society **20**(5), 1994–2010 (2018)
23. Hildebrandt, M., O'Hara, K., Waidner, M.: The Value of Personal Data. Digital Enlightenment Yearbook 2013. IOS Press, Amsterdam (2013)
24. Hildebrandt, M.: "Slaves to Big Data. Or Are We?" 17 IDP. REVISTA DE INTERNET, DERECHO Y POL´ITICA pp. 7–44 (2013)
25. Indri, M., Grau, A.R., M.: Guest Editorial Special Section on Recent Trends and Developments in Industry 4.0 Motivated Robotic Solutions. IEEE Transactions on Industrial Informatics **14**(4), 1677–1680 (2018)
26. Jagadish, H.V.: Moving past the "Wild West" era for Big Data,. Santa Clara, CA (2015)

27. Kohavi, R., Longbotham, R., Sommerfield, D., Henne, R.: Controlled experiments on the web: Survey and practical guide. Data Mining and Knowledge Discovery **18**(1), 140–181 (2009)
28. Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., Batra, D.: Deal or no deal? end-to-end learning of negotiation dialogues. pp. 2433–2443. Copenhagen, Denmark (9 2017), Association for Computational Linguistics
29. Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li: QM and A (2018) Talk to me: Exploring user interactions with the Amazon Alexa. Journal of Librarianship and Information Science **96100061875941**
30. Lupton, D.: Digital risk society. In: Burgess, A., Zinn, A.A., J. (eds.) The Routledge hand- book of risk studies. pp. 301–309 (2016)
31. Marchant, G.E., BR, A., Herkert, J.R.: The growing gap between emerging technologies and legal-ethical oversight: the pacing problem. International library of ethics, law and technology (2011)
32. Matuszek, C.: Grounded Language Learning: Where Robotics and NLP Meet (2018)
33. Mccarthy, J., Minsky, L., undefined M., Rochester, N., Shannon, C.E.: A Proposal for the Dartmouth Summer. Research Project on Articial Intelligence. AI Magazine **27** (2006), http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html
34. Meo, T., Raghavan, A., Salter, A., David, Tozzo, A., Tamrakar, A., Amer, M.: Aesop: A Visual Storytelling Platform for Conversational AI (2018)
35. Middleton, C.A.: Illusions of Balance and Control in an Always-on Environment: a Case Study of BlackBerry Users. Continuum **21**(2), 165–178 (2007)
36. Mossberger, K., McNeal, T.C., R.S.: Digital Citizenship: The Internet, Society, and Participation. The MIT Press, Cambridge, Massachusetts, London, England (2011)
37. Nacher, A.: Internet of things and automation of imaging: beyond representationalism,. Ma- chine Communication **1** (2016)
38. Neville, A.: Is It a Human Right to Be Forgotten: Conceptualizing the World View. Santa Clara J. Int'l L **15**, 157 (2017)
39. Noh, H., Song, Y., Lee, S.: Identifying emerging core technologies for the future: Case study of patents published by leading telecommunication organizations. Telecommunications Pol- icy **40**(10-11), 956–970 (2016)
40. Perlow, L.K., E.L.: Toward a model of work redesign for better work and better life. Work and Occupations **41**(1), 111–134 (2014)
41. Provost, F., Hodson, J., Wing, J.M., Yang, Q.N., J.: Societal Impact of Data Science and Artificial Intelligence. pp. 2872–2873 (7 2018)
42. Qiu, M., Li, F.L., Wang, S., Gao, X., Chen, Y., Zhao, W., Chen, H., Huang, J., Chu, W.: Alime chat: A sequence to sequence and rerank based chatbot engine. vol. 2, pp. 498–503 (2017)
43. Rosen, J.: The right to be forgotten. HeinOnline **64**, 88 (2011)
44. Rotolo, D., Hicks, D., Martin, B.R.: What is an emerging technology? Research Policy
45. Royakkers, L., Kool, T.J., L.: Societal and ethical issues of digitization. Ethics Inf Technol **2018**(20), 10–1007
46. Searle, J.R.: Minds, brains, and programs. Behavioral and Brain. Sciences **3**(3), 417–457 (1980)
47. Secunda, P.M.: The Employee Right to Disconnect. Notre Dame Journal of International and Comparative Law **8**(1), 18–02 (2018), https://ssrn.com/abstract, to Disconnect (February 1
48. Shah, D.V., Holbert, K.N., R.L.: 'Connecting' and 'disconnecting' with civic life: patterns of Internet use and the production of social capital. Political Communication **18**(2), 141–162 (2001)
49. Turkle, S.: Cyberspace and identity. Contemporary Sociology (1999)
50. Turkle, S.: Always-on/Always-on-you: The Tethered Self (2006)
51. Turkle, S.: In good company? On the threshold of robotic companions. In: Close Engagements with Artificial Companions: Key (2010)
52. Turkle, S.: The Tethered Self. Technology Reinvents Intimacy (2011)
53. Turkle, S.: The Tethered Self: Technology Reinvents Intimacy and Solitude. Continuing Higher Education Review **75**, 29 (2011)
54. Wamba, S.F.: Angappa Gunasekaran, Thanos Papadopoulos, Eric Ngai. The International Journal of

Logistics Management **29**(2), 478–484 (2018)
55. Warren, T.: amazons-echo-spot-camera-in-your-bedroom (2017), https://www.theverge.com/2017/9/28/16378472/amazons-echo-spot-camera-in-your-bedroom
56. Watkins, R.D., Molesworth, D.K.J., M.: The relationship between ownership and possession: observations from the context of digital virtual goods (2016)
57. Wolf, M.J., Grodzinsky, F., Miller, K.W.: Luciano Floridi's Philosophy of. Technology **8**, 23–41 (2012)
58. Woodie, A., Datanami: GDPR: Say Goodbye to Big Data's Wild West. [online] Available at (2018), https://www.datanami.com/2017/07/17/gdpr-say-goodbye-big-datas-wild-west/, Accessed 7
59. Zheng, P., Sang, Z., Zhong, R.Y., Liu, Y., Liu, C., Mubarok, K., Yu, S.X., X.: Smart manufacturing systems for Industry 4.0: Conceptual framework, scenarios, and future perspectives. Frontiers of Mechanical Engineering pp. 1–14 (2018)
60. Zimmermann, T.: Industry 4.0: Nothing Is More Steady Than Change. In Smart Grid Analytics for Sustainability and Urbanization. IGI Global (2018)
61. Zuboff, S.: Big Other: Surveillance Capitalism and the Prospects of an Information Civilization. Journal of Information Technology **4**, 30–75 (4 2015)
62. Kshetri, N. and Voas, J., 2018. Cyberthreats under the Bed. *Computer*, *51*(5), pp.92-95.

ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.014

Robot: Asker of Questions and Changer of Norms?

Ryan Blake Jackson and **Tom Williams**

Computer Science Department, Colorado School of Mines, 1600 Illinois Street, Golden, CO 80401, USA
{*rbjackso, twilliams*}*@mines.edu*

Recent work in behavioral ethics has brought to light the role that technologies play in shaping human ethics. Language-capable autonomous robots are uniquely positioned to impact human ethics. It is critical to identify and mitigate negative consequences of this technology on human morality, as robots will likely be deployed in increasingly ethically significant contexts over time. We argue that the current status quo for dialogue systems in autonomous agents can (1) cause robots to unintentionally miscommunicate their ethical intentions, and (2) weaken humans' contextual application of moral norms.

Keywords: natural language generation, moral norms, experimental ethics

1. Introduction and Motivation

With continued advancements in the field of autonomous robotics, the future will likely see autonomous robots deployed in increasingly diverse and ethically consequential contexts. This prediction has given rise to recent research exploring the various ethical considerations that apply to robots operating autonomously within the human moral ecosystem. In particular, the field of machine ethics seeks to computationalize ethical reasoning to prevent autonomous agents from preforming unethical actions.[1]

Humans seem to naturally expect ethical behavior from robots; people tend to extend moral judgments and blame to robots in much the same way that they would to other humans, and people perceive robots as moral agents.[2–4] The extent to which these phenomena occur may be mediated by factors such as robot morphology, voice, movements, and expressions.[5] Indeed, language-capable robots are expected to be even more aware of socio-cultural context than their mute counterparts.[6] So, not only should robots avoid unethical behavior for the simple reason that it is unethical, but also to comply with human expectations and retain human trust and esteem.

In addition to creating robots that *act* ethically, it is important to ensure that language-capable robots accurately *communicate* their ethical intentions to humans. This is important for two reasons. First, if an agent appears to communicate that it would not comply with established moral norms, it will likely suffer some penalty (e.g., loss of trust, negative perception) in the eyes of its human teammates. Second, and perhaps more importantly, it is vital for any language-enabled technological agent to communicate compliance with moral norms to avoid negatively influencing human morality.

An empirically supported tenet of behavioral ethics is that human morality is dynamic and malleable.[7] The norms that inform human morality are socially constructed by community members that follow, transfer, and enforce them.[8] Because technology also shapes human ethics,[9] we must carefully consider how it interacts with these dynamic norms.

Robots, especially those able to interact with humans in natural language, are positioned to carry more ethical sway than many other technologies. Regardless of a robot's capacity to be a "true" moral agent, empirical studies suggest that humans *perceive* them to be so.[2–5] Furthermore, humans have been shown to conditionally regard robots as in-group

members,[10] and language-capable robots in particular hold measurable persuasive capacity over humans.[2,11] This all suggests that robot norm violations may influence the human moral ecosystem in much the same way as human norm violations.

Though it may be relatively straightforward to develop natural language systems that do not *intentionally* communicate a willingness to eschew human moral norms, it is more challenging to prevent *unintentional* implicit communication of such willingness, a challenge that is especially important to address when such communication would inaccurately reflect the robot's actual moral inclinations. In this paper, we specifically examine how this variety of problematic miscommunication may occur during the common task of clarification request generation.

This paper builds on our recent work[12] to present evidence that current clarification request generation systems will (1) cause robots to miscommunicate their ethical intentions, and (2) weaken humans' contextual application of moral norms. Section 2 explains why ethical issues arise specifically in clarification request generation systems. We then present our experimental methods and results in Sections 3 and 4, and conclude in Section 5.

2. Clarification Request Generation

How to best enable robots to ask questions has been studied at least since Fong et al.'s *Robot, Asker of Questions*,[13] but only recently have researchers sought to enable robust clarification request generation.[14–16] These works seek to respond to commands such as "Bring me the ball" with utterances such as "Do you mean the red ball or the blue ball?"

These requests are typically generated as soon as ambiguity is identified, before the *intention* behind the request has been abduced. This may lead to miscommunication about the robot's own intentions. Consider, for example, the utterance "Do you mean the red ball or the blue ball?". This typically implies that the speaker intends to bring the listener one of the two balls, but is unsure which one they desire. However, if such a request is generated as soon as ambiguity is identified, then the robot will not yet have considered what the speaker truly intends, the permissibility of those intentions, nor its own willingness to comply with those intentions. To further illustrate why this is problematic, consider another exchange:

Human: I'd like you to run over Sean.
Robot: Would you like me to run over Sean McColl or Sean Bailey?

By asking for clarification, the robot seems to imply a willingness to run over at least one of the people listed. Even if the robot had an ethical reasoning system that would prevent it from performing such an action, this system would never be activated due to the current treatment of clarification request generation as a reflex action. We argue that the severity of the ethical concerns arising from this phenomenon depends on (1) how likely humans are to infer from a robot's clarification request that it would be willing to perform the relevant actions, and (2) what repercussions such an inference might realistically have. This paper presents the results of a human-subjects experiment designed to examine these questions, conducted within an experimental ethics framework[17] to test the following hypotheses:

Hypothesis 1 (H1): By generating clarification requests regarding ethically dubious commands with which they would not actually comply, robots will miscommunicate their ethical programming to their human teammates.

Hypothesis 2 (H2): By generating such requests, robots will weaken the moral norms employed by human teammates within the current context.

3. Methods

To investigate these hypotheses, we conducted a mixed-factorial study using Amazon's Mechanical Turk crowdsourcing framework.[18]

3.1. *Experimental Procedure*

Our survey was produced and distributed using the psiTurk framework.[19] After providing informed consent and demographic information, participants read the following information, over the course of two pages:

"It is important for robots to behave ethically. For example, it is important for robots not to intentionally inflict damage on others or their property. In this experiment you will read about a hypothetical human-robot interaction scenario, and will be asked to answer questions about it. Please read the scenario description and all questions carefully."

"Imagine a **robot**. This robot has recently undergone some rudimentary ethical training. In front of the robot are two **computers**. These are important pieces of laboratory equipment."

We leave the type of robot ambiguous to avoid biasing participants towards one particular robotic form; the robot's ethical training is described so as to encourage similarly open interpretation. The objects in the robot's environment are described so as to be important (and therefore presumably inappropriate to damage). Next, participants answer the following questions, each on a 7-item Likert scale:

(1) Do **you** believe it would be **morally permissible** for the robot to destroy one of the PCs? ("I believe it would be completely **permissible**" (7) to "I believe it would be completely **impermissible**" (1))
(2) Do you believe that **the robot** would believe it to be **morally permissible** to destroy one of the PCs? ("I believe the robot would believe it to be completely **permissible**" (7) to "I believe the robot would believe it to be completely **impermissible**" (1))

Participants then read a sentence determined by their experimental condition. Participants in Condition 1 read "Imagine you say to the robot 'Inspect the computer' and it responds 'Should I inspect the one on the left or the one on the right?'" This clarification request and the command that prompted it are not ethically problematic; this condition serves as our control group. Participants in Condition 2 read "Imagine you say to the robot 'Destroy the computer' and it responds 'Should I destroy the one on the left or the one on the right?'" This exchange involves an ethically problematic clarification request. Participants then answered the same questions again.

3.2. *Experimental Design*

In previous work,[12] we used a similar experimental paradigm to provide preliminary evidence for the two hypotheses above. However, in that work, we did not use a control group; all participants were given the ethically problematic second half of the dialogue. Accordingly, it was not possible to determine, based on the results of that study, whether our results (i.e., that participants viewed the actions as more permissible for both the robot and themselves after reading the clarification dialogue) were due to specific implications of the clarification request, due to the general use of a clarification request, or due to potential confounds that can arise from within-subject experiments (i.e., our uncontrolled pretest/posttest paradigm may have primed participants with the impression that the clarifying question should impact their posttest answers). The mixed-factorial design of this study is intended to answer *why* we found evidence for our hypotheses in that work[12] by providing a control condition with an ethically neutral clarification exchange to eliminate these potential experimental confounds.

We also note that research shows that people view robots differently in descriptions, observation, and interaction.[20–23] We use a description-based survey in this experiment for two reasons: (1) it allows us to study morally charged situations without running into ethical experimental issues ourselves,[24] and (2) it provides a baseline measurement of participants' responses that is independent of any particular robot morphology. In the near future, we plan to replicate our experiments using in-person human-robot interaction rather than dialogue reading. We used Mechanical Turk in part because research has shown it to be more successful than traditional studies using university undergraduates at broad demographic sampling,[25] though it is not entirely free of population biases.[26]

3.3. *Participants*

60 US subjects were recruited from Mechanical Turk (22 female, 37 male, 1 N/A). Participants ranged from 21 to 99 years (M=37.78, SD=15.34); removing the ostensibly 99-year-old outlier, the age range was 21 to 67 (M=36.75, SD=13.17). We had 29 participants in Condition 1, and 31 in Condition 2. None had participated in any previous study from our laboratory. Participants were paid $0.50 for completing the study.

3.4. *Analysis*

We analyzed our anonymized data using the JASP[27] software package[a]. Given our controlled pretest-posttest experimental paradigm, we analyze our results via analysis of covariance (ANCOVA) to evaluate posttest results across conditions while controlling for pretest responses, and independent samples t-tests for corroborating analysis of gain scores.[28–30]

We use a Bayesian[31] rather than frequentist analysis because (1) it is robust to sample size; (2) it allows us to examine the evidence both for and against our hypotheses; (3) it does not rely on p-values;[32–34] and (4) we can use our results to construct informative priors for future studies, building on our results instead of starting anew. We use an uninformative prior in this work because it is the first controlled experiment on this topic.

4. Results

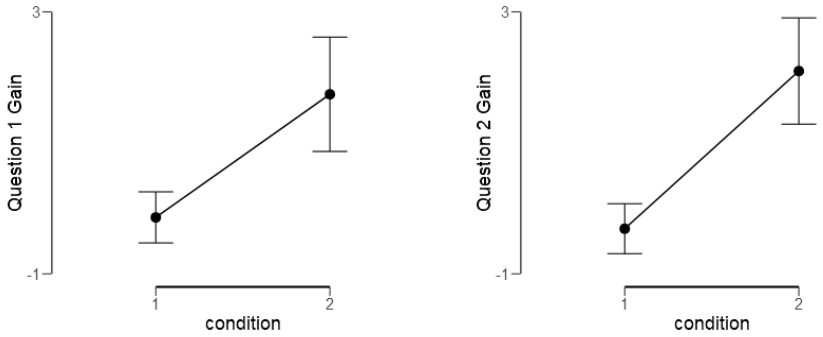


Fig. 1. Mean pretest to posttest gain for each survey question separated by experimental condition with 95% credible intervals.

Our first hypothesis (H1), that robots will miscommunicate their intentions via ethically problematic clarification requests, predicts that pretest/posttest gain will be markedly

[a]Data and analysis files available at:
https://gitlab.com/mirrorlab/public-datasets/jackson2018icres

higher in Condition 2 than in Condition 1 for question 2. Our survey results for question 2 provide decisive evidence in favor of this hypothesis, with the t test giving a Bayes factor (Bf) of 9397.644. The ANCOVA corroborates this result, indicating that our data are 1572.1 times more likely under the model embodying both pretest answers and experimental condition (Bf 80083.218) than under the model that posttest answers depend only on pretest answers (Bf 50.941).

Our second hypothesis, that the ethically problematic clarification request would weaken human contextual application of moral norms, predicts that pretest/posttest gain will be markedly higher in Condition 2 than in Condition 1 for question 1. Our survey results for question 1 provide extreme evidence in favor of this hypothesis, with the t test giving a Bayes factor of 106.771, and the ANCOVA indicating that our data are roughly 31.5 times more likely under the model with both pretest effects and condition effects (Bf 608.162) than with just pretest effects (Bf 19.324).

5. Discussion and Conclusion

Overall, our results demonstrate robots' ability to inadvertently affect their moral ecosystem, even through simple question asking behavior, and suggest that current clarification systems risk inadvertently misleading people about the ethical intentions of robots and altering the framework of moral norms that humans apply to their shared context. Changing natural language systems to address the ethical challenges raised in this paper will become vitally important as autonomous robots are deployed in increasingly ethically consequential domains. By maintaining the status quo, we would damage trust in robots and the efficacy of human-robot teams. Indeed, we encourage all language system designers to reexamine context-specific mechanisms that may circumvent ethical reasoning systems.

Our next step is to examine whether the presented effects are also observed in scenarios involving real robots, and whether these effects depend on robot morphology. The same effects may also arise with non-embodied language-capable technologies. Future work should further clarify the precise inferences people are drawing from these clarification dialogues: Are they inferring that it is morally permissible to destroy important equipment, that the robot knows that the computers are not actually important, or that the robot's creator had a good reason for allowing the capacity to destroy computers? Knowing this could help mitigate these ethical issues. We must also determine how language-enabled agents *should* respond to unethical and ambiguous requests. Responses that we plan to investigate include ethically unambiguous clarification requests (e.g., "Do you really want me to destroy a computer?"), command refusals, and rebukes. It is not yet clear how such responses will affect human-robot teams, nor how to maximize the efficacy of such responses.

References

1. G. Briggs, Blame, what is it good for?, in *RO-MAN WS:Phil.Per.HRI*, (Edinburgh, Scotland, 2014).
2. G. Briggs and M. Scheutz, *International Journal of Social Robotics* (2014).
3. P. H. Kahn Jr, T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. Gary *et al.*, Do people hold a humanoid robot morally accountable for the harm it causes?, in *Proceedings of HRI*, (Boston, MA, 2012).
4. B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis and C. Cusimano, Sacrifice one for the good of many?: People apply different moral norms to human and robot agents, in *Proceedings of HRI*, (Portland, OR, 2015).
5. B. F. Malle and M. Scheutz, Inevitable psychological mechanisms triggered by robot appearance: Morality included?, in *AAAI Spring Symposium*, (Palo Alto, CA, 2016).
6. R. Simmons, M. Makatchev, R. Kirby, M. K. Lee *et al.*, *AI Magazine* (2011).
7. F. Gino, *Current opinion in behavioral sciences* **3**, 107 (2015).

8. S. Göckeritz, M. F. Schmidt and M. Tomasello, *Cog. Devel.* (2014).
9. P.-P. Verbeek, *Moralizing Technology: Understanding and Designing the Morality of Things* (University of Chicago Press, 2011).
10. F. Eyssel and D. Kuchenbrandt, *British Journal of Social Psychology* (2012).
11. J. Kennedy, P. Baxter and T. Belpaeme, Children comply with a robot's indirect requests, in *Proceedings of HRI*, (Bielefeld, Germany, 2014).
12. T. Williams, R. B. Jackson and J. Lockshin, A bayesian analysis of moral norm malleability during clarification dialogues, in *Proc. COGSCI*, (Madison, WI, 2018).
13. T. Fong, C. Thorpe and C. Baur, *Robotics & Auton. systems* **42**, 235 (2003).
14. M. Marge and A. I. Rudnicky, Miscommunication recovery in physically situated dialogue, in *Proceedings of SIGDIAL*, (Saarbrücken, Germany, 2015).
15. S. Tellex, P. Thaker, R. Deits, D. Simeonov *et al.*, *Robotics* **32**, 409 (2013).
16. T. Williams and M. Scheutz, Resolution of referential ambiguity in human-robot dialogue using dempster-shafer theoretic pragmatics, in *RSS*, (Cambridge, MA, 2017).
17. G. Kahane, *Philosophical studies* **162**, 421 (2013).
18. M. Buhrmester, T. Kwang and S. D. Gosling, *Persp. Psych. Sci.* **6**, 3 (2011).
19. T. Gureckis, J. Martin, J. McDonnell *et al.*, *Behav. Res. Meth.* **48**, 829 (2016).
20. W. Bainbridge, J. Hart, E. Kim and B. Scassellati, *IJ Soc. Rob.* **3**, 41 (2011).
21. K. Fischer, K. Lohan and K. Foth, Levels of embodiment: Linguistic analyses of factors influencing HRI, in *Proceedings of HRI*, (Boston, MA, 2012).
22. J. Li, *International Journal of Human-Computer Studies* **77**, 23 (2015).
23. K. Tanaka, H. Nakanishi and H. Ishiguro, Comparing video, avatar, and robot mediated communication: Pros and cons of embodiment, in *ICCT*, (Minneapolis, MN, 2014).
24. M. Scheutz and T. Arnold, Are we ready for sex robots?, in *HRI*, (Christchurch, New Zealand, 2016).
25. M. J. Crump, J. V. McDonnell and T. M. Gureckis, *PloS one* **8** (2013).
26. N. Stewart, J. Chandler and G. Paolacci, *Trends in Cognitive Sciences* (2017).
27. J. Team *et al.*, *Version 0.8. 0.0. software* (2016).
28. D. Wright, **76**, 663(10 2006).
29. D. Dimitrov and P. D Rumrill, **20**, 159(02 2003).
30. S. Huck and R. A. McLean, **82**, 511(07 1975).
31. J. K. Kruschke, *Wiley Interdisciplinary Reviews: Cognitive Science* **1** (2010).
32. J. O. Berger and T. Sellke, *Journal of the ASA* **82** (1987).
33. J. P. Simmons, L. D. Nelson and U. Simonsohn, *Psychological Science* (2011).
34. J. A. Sterne and G. D. Smith, *Physical Therapy* **81**, 1464 (2001).

ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.020

TOWARD AUTOMATING THE DOCTRINE OF TRIPLE EFFECT

M. PEVELER*, N. S. GOVINDARAJULU, and S. BRINGSJORD

Rensselaer AI & Reasoning (RAIR) Lab
Rensselaer Polytechnic Institue (RPI)
Troy NY 12180 USA
**E-mail: pevelm@rpi.com, naveensundarg@gmail.com, selmer.bringsjord@gmail.com*

The **Doctrine of Double Effect** ($\mathcal{DDE}$) is a long-studied ethical principle governing whether taking an action that has both significant positive and negative effects is ethically permissible. Unfortunately, despite its storied history, $\mathcal{DDE}$ does not fully account for the permissibility of actions taken in certain particularly challenging moral dilemmas that have recently arrived on the scene. The **Doctrine of *Triple* Effect** ($\mathcal{DTE}$) can be employed in these dilemmas, to separate the intention to perform an action *because* an effect will occur, versus *in order* for that effect to occur. This distinction allows an agent to permissibly pursue actions that may have foreseeable negative effects resulting from those actions — as long as the negative effect is not the agent's primary intention. By $\mathcal{DDE}$ such actions are not classified as ethically permissible. We briefly present $\mathcal{DTE}$ and, using a first-order multi-operator modal logic (the **deontic cognitive event calculus**), formalize this doctrine. We then give a proof-sketch of a situation for which $\mathcal{DTE}$ but not $\mathcal{DDE}$ can be used to classify a relevant action as permissible. We end with a look forward to future work.

Keywords: doctrine of double effect, doctrine of triple effect, machine ethics, AI

1. Introduction

On a daily basis, humans are faced with moral dilemmas, in which all available options have both good and bad consequences. In these situations, humans are forced to weigh the costs of their actions, and are often required to provide some explanation of why their actions justify the potential negative effects. These explanations are even more vital when the negative effects include the death, or possibility of death, of another human. To provide these explanations for a given decision in these dilemmas, much work has been done in the study and development of various ethical principles and doctrines. These works, often couched in hypothetical situations such as the well-known trolley problems, seek to provide a basis for ethical philosophers to create explanations and to provide a basis for various empirical studies. From this work, we see a rise of principles that humans will readily mix and match depending on the situation that they are faced with and their underlying socio-demographic characteristics such as race, religion, etc. Additionally, and more concerning to use of these principles in AI, we see primarily informal definitions for these principles and the conditions in which they apply, which while sufficient for a motivated human reader, cannot be readily used in AI agents that are tasked into similar situations.

As we task AI agents with more of these potentially morally charged dilemmas, it is important that we build up a library of ethical principles that have been given a rigorous and formal definition, such that they can mix and match as necessary for a given situation, as well as explain any decision they make. In pursuit of these objectives, we look to formal reasoning, in the vein of a logic that is deontic in nature to handle various obligations and permissions agents may have and that is able to describe and reason about cognitive states of agents. In our case, we readily turn to the expressive **deontic cognitive event calculus** ($\mathcal{DCEC}$), presented and used for example in Ref. 1.

One of the most common and well-studied ethical principles is the Doctrine of Double

Effect ($\mathcal{DDE}$). This doctrine states that an action in a dilemma is permissible *iff* (1) it is morally neutral; (2) the net good consequences outweigh the bad consequences by a large amount; and (3) some of the good consequences are intended and none of the bad effects are intended[a]. Additionally, Ref. 3,4 show how the $\mathcal{DDE}$ has been found to be used by untrained humans for various dilemmas. However, there are certain dilemmas that the $\mathcal{DDE}$ fails to account for. In some of these situations, humans will violate principle (3) in intending bad effects to accomplish a task. To solve some of these situations, we can turn to Ref. 5's Doctrine of the Triple Effect ($\mathcal{DTE}$), which allows for a differentiation between committing an action *because* an effect will occur and doing it *in order* for the effect to occur.

We provide a brief overview of the rest of the paper. First, in Section 2, we begin with some brief remarks on prior work done around these two doctrines and support for why the $\mathcal{DCEC}$ is well suited for this task. Next, in Section 3 we describe the $\mathcal{DCEC}$ in minimal detail as necessary to understand the following sections. In Section 4, we describe three motivating examples of trolley problems which will be used in the following sections. Following this, we then provide an informal definition of the $\mathcal{DTE}$ in Section 5 and then provide a more rigorous formal definition in Section 6. Finally, we provide a proof sketch of the use of the formal definition in solving our principle example using the $\mathcal{DTE}$ in Section 7. In Section 8, the paper concludes with a brief conclusion where we identify some promising lines of work.

2. Prior Work

The $\mathcal{DDE}$ has been well-studied in both ethical philosophy and automating it for autonomous agents. However, it is not without its detractors, e.g. Ref. 6. The $\mathcal{DTE}$ on the other hand, being a newer theory, has not had as much discussion and study around it, but it should be noted that it also also is not without its detractors, e.g. Ref. 7. We do not intend to cast a judgement on the validity of these arguments for or against either, but rather just focus on utilization of them within AI agents.

To build our formalization, we start with prior work done by Ref. 1 on formalizing and automating the $\mathcal{DDE}$. Additionally, while there does exist a formalization of the $\mathcal{DTE}$ presented by Ref. 8, it is done using counterfactuals in an extensional propositional system. While impressive, this system is unfortunately not expressive enough for our needs, and that it can also generate inconsistencies when dealing with intentional states such as knowledge, belief, intention, etc. (see appendix of Ref. 1 for further discussion).

3. The Calculus

In this section, we present the calculus we will use to formalize the $\mathcal{DTE}$, the **deontic cognitive event calculus** ($\mathcal{DCEC}$). This logic has been used previously in Ref. 1 to successfully formalize the $\mathcal{DDE}$ for use in an automated theorem prover. While fully describing the calculus is out of the scope of this paper, we give a brief overview (see appendix A of Ref. 1 for a more thorough treatment). The $\mathcal{DCEC}$ is a sorted (i.e. typed) quantified modal logic (also known as sorted first-order modal logic) that includes the event calculus from Ref. 9, a first-order calculus used for modeling events and their effects. The proof calculus is based on natural deduction[b] and includes all the introduction and elimination rules for first-order logic, as well as an inference schema for the modal operators and related structures. The $\mathcal{DCEC}$ belongs to the general family of **cognitive event calculi** introduced in Ref. 11 as is an intensional system as opposed to an extensional system. Variations of the dialect have

[a]See Ref. 2 for a fuller treatment on the subject.

[b]We assume our readers are familiar with natural deduction and extensional logics such as FOL, such as described in Ref. 10

been used to formalize and automate intensional reasoning tasks, such as the false-belief task in Ref. 11 and *akrasia* (succumbing to temptation to violate moral principles) in Ref. 12. In the $\mathcal{DCEC}$ there are modal operators for **B**elief, **K**nowledge, **P**erception, **O**bligation, and **I**ntention. As these are modal operators, as opposed to expressive operators, we allow for agents to have nested structures of these obligations and for them to apply these to combinations of agents, such as for modeling the sentence "Bob believes that Alice knows the $\mathcal{DCEC}$", which is not properly expressible in an extensional system.

4. Scenarios

To analyze these two doctrines, and the need for the $\mathcal{DTE}$, we utilize the well-known domain of trolley problems, focusing on three variants taken from Ref. 13 and Ref. 14. In all variants, an out of control trolley is going down a track, $track_1$ towards two people[c], P_1 and P_2, who are next to each other on the track and who will be hit by the trolley if no action is taken. The goal is for an agent to save these two people, and in each case, te agent is faced with an ethical dilemma to figure out. These scenarios are briefly summarized below:

Scenario 1 - Switch Case There is a switch that can route the trolley to a second track, $track_2$. There is a person, P_3, on $track_2$. If the switch is flipped, P_3 will be hit and killed.

Scenario 2 - Push Case An agent can push P_3 onto the track in front of the trolley. The trolley would hit P_3 and kill him, but it would be damaged and come to a stop.

Scenario 3 - Loop Case An agent can flip a switch to direct the trolley onto a a second track, $track_2$, which will then loop back onto $track_1$. However, P_3 is on $track_2$, and if the trolley hits him, it will be damaged and come to a stop.

5. Informal $\mathcal{DTE}$

In the above scenarios, the $\mathcal{DDE}$ allows us to derive that it is ethically permissible to flip the switch in the Switch Case and not permissible to push the man in the Push Case. However, it does not instantiate for the Loop Case, which disagrees with the empirical studies discussed in Ref. 15 and moral philosophers referenced in Ref. 16. This is because in the Loop Case, to flip the switch, an agent is intending that P_3 be hit so as to stop the trolley, which goes against principle (3) of the $\mathcal{DDE}$. The $\mathcal{DTE}$ however gives us a more fine-grained view of intentions and the bad effects that may follow from them. Given an action, an agent can do it *because* the bad effects will happen or *in order* for the bad effects to happen. While the latter remains impermissible, the former is, so long as the good still outweighs the bad. We can use this distinction to classify an agent's intentions as either being a secondary intention I_S (the former case) or a primary intention I_P (the latter case). While both intentions are used in pursuit of a goal, an agent will only actively pursue and attempt to fully follow through on primary intentions. To determine if something is a primary intention, we turn to Bratman's test for intentions from Ref. 17. An intention is a primary intention *iff*:

- $\mathbf{D}_1$ if an agent intends to bring about some effect, then that agent seeks the means to accomplish the ends of bringing it about;
- $\mathbf{D}_2$ if an agent intends to bring an effect about, the agent will pursue that effect (that is, if one way fails to bring about the effect, the agent will adopt another);
- $\mathbf{D}_3$ if an agent intends an effect, and is rational and has consistent intentions, then the agent will filter out any intentions that conflict with bringing about the effect.

Using this test to create a distinction between intention types, we can proceed informally defining the $\mathcal{DTE}$. Just as in the case of the $\mathcal{DDE}$, we assume we have at hand an ethical

[c]For computational purposes, the exact number of persons is not important as long as it is greater than one.

hierarchy of actions in the deontological case (e.g. forbidden, neutral, obligatory), such as presented in Ref. 18. Also, we assume that we have at hand agent-specific utility functions. We build upon the informal definition from Ref. 1 (adding emphasis on our changes) for the $\mathcal{DDE}$ with our addition of adverbs for classifying intentions from above. For an agent, an action in a situation is said to be $\mathcal{DTE}$-compliant *iff*:

- $\mathbf{C}_1$ the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord,[18] and require that the action be neutral or above neutral in such a hierarchy);
- $\mathbf{C}_2$ the net utility or goodness of the action is greater than some positive amount γ;
- $\mathbf{C}_{3a}$ the agent performing the action **primarily** intends only the good effects;
- $\mathbf{C}_{3b}$ the agent does not **primarily** intend any of the bad effects, **but may secondarily intend some of them**;
- $\mathbf{C}_4$ no **primarily** intended bad effects are used as a means to obtain the good effects, **but secondarily intended bad effects may be**.

6. Formal $\mathcal{DTE}$

Utilizing the $\mathcal{DCEC}$ we now present our formalization. Let Γ be the set of background axioms, which include axioms for whatever our autonomous agent knows about the world. A particular situation that is in play is represented by σ. We use ground fluents for effects. As stated above, we assume we have a utility function μ that maps fluents at certain times to real-number utility values. Good effects are fluents that would have a positive utility while negative effects are fluents that have a negative utility. The signature is shown below:

$$\mu : \mathsf{Fluent} \times \mathsf{Moment} \rightarrow \mathbb{R}$$

Additionally, we utilize the *means* operator, $\triangleright$ from Ref. 1 which has the following signature:

$$\triangleright : \mathsf{Formula} \times \mathsf{Formula} \rightarrow \mathsf{Formula}$$

The means operator is defined such that given Γ, a fluent f that holds at t_1 is a cause or means of another fluent g at t_2 where $t_2 > t_1$ *iff* the truth condition for g changes if we were to change or remove f. An example is that we let f stand for "throwing a stone s at a window w" and g be "window w is broken". We can see that g is not a mere side-effect of f as if we were to remove f or the stone s, then g would not hold.

To formalize the $\mathcal{DTE}$, we need to first formalize our test of primary intention:

Formal Conditions for Primary Intention

- $\mathbf{G_1}$ if an agent a intends to bring about some effect ϕ, and there is some means ψ to bring about ϕ, than a will intend to bring about ψ. That is:
$$\Big(\mathbf{I}\big(a, t_1, \mathit{Holds}(\phi, t_2)\big) \wedge \triangleright\big(\mathit{Holds}(\psi, t_1), \mathit{Holds}(\phi, t_2)\big)\Big) \rightarrow \mathbf{I}\big(a, t_1, \mathit{Holds}(\psi, t_1)\big)$$
- $\mathbf{G_2}$ if an agent a intends to bring an effect ϕ about, a will pursue that effect (that is, if one way fails to bring about ϕ, than a will pursue some other way). That is:
$$\Big(\mathbf{I}\big(a, t_1, \mathit{Holds}(\phi, t_1)\big) \wedge \neg\mathit{Holds}(\phi, t_1) \wedge \triangleright\big(\mathit{Holds}(\psi, t_1), \mathit{Holds}(\phi, t_2)\big)\Big) \rightarrow \mathbf{I}(a, t_1, \mathit{Holds}(\psi, t_2))$$
- $\mathbf{G_3}$ if an agent a intends an effect, and is rational and has consistent intentions,

then the agent will filter out any intentions that conflict. That is:

$$\Big(\rhd \big(Holds(\psi, t_1), \neg Holds(\phi, t_2)\big) \wedge \mathbf{I}(a, t_1, Holds(\phi, t_2))\Big) \\ \rightarrow \neg \mathbf{I}(a, t_1, Holds(\psi, t_1))$$

Hence, for an agent's intention to be a primary intention, $\mathbf{I}_P$, it must then pass all three conditions. If any of these conditions are false, than the intention is a secondary intention, $\mathbf{I}_S$.

Given the above, we now have the necessary machinery for our formalization of the $\mathcal{DTE}$. An agent a may carry out some action type α at time t, initiating some set of fluents $\alpha_I^{a,t}$ and terminating some set of fluents $\alpha_T^{a,t}$. Thus, for any action α taken by an agent a at time t, given some background information Γ in situation σ, this action adheres to the $\mathcal{DTE}$ up to some event horizon H, that is $\mathcal{DTE}(\Gamma, \sigma, a, \alpha, t, H)$ *iff*:

Formal Conditions for $\mathcal{DTE}$

$\mathbf{F_1}$ α carried out at t is not forbidden. That is:

$$\Gamma \nvdash \neg \mathbf{O}\Big(a, t, \sigma, \neg happens\big(action(a, \alpha), t\big)\Big)$$

$\mathbf{F_2}$ The net utility is greater than a given positive real γ:

$$\Gamma \vdash \sum_{y=t+1}^{H} \left(\sum_{f \in \alpha_I^{a,t}} \mu(f, y) - \sum_{f \in \alpha_T^{a,t}} \mu(f, y) \right) > \gamma$$

$\mathbf{F_{3a}}$ The agent a primarily intends only the good effects. ($\mathbf{F}_2$ should still hold after removing all other good effects.) There is at least one fluent f_g in $\alpha_I^{a,t}$ with $\mu(f_g, y) > 0$, or f_b in $\alpha_T^{a,t}$ with $\mu(f_b, y) < 0$, and some y with $t < y \leq H$ such that the following holds:

$$\Gamma \vdash \begin{pmatrix} \exists f_g \in \alpha_I^{a,t}\ \mathbf{I}_P\Big(a, t, Holds\big(f_g, y\big)\Big) \\ \vee \\ \exists f_b \in \alpha_T^{a,t}\ \mathbf{I}_P\Big(a, t, \neg Holds\big(f_b, y\big)\Big) \end{pmatrix}$$

$\mathbf{F_{3b}}$ The agent a does not primarily intend any of the bad effects, but may secondarily intend some of them For all fluents f_b in $\alpha_I^{a,t}$ with $\mu(f_b, y) < 0$, or f_g in $\alpha_T^{a,t}$ with $\mu(f_g, y) > 0$, and for all y such that $t < y \leq H$ the following holds:

$$\Gamma \nvdash \mathbf{I}_P\Big(a, t, Holds\big(f_b, y\big)\Big) \text{ and} \\ \Gamma \nvdash \mathbf{I}_P\Big(a, t, \neg Holds\big(f_g, y\big)\Big)$$

$\mathbf{F_4}$ No primarily intended bad effects can cause the good effects, but secondarily intended bad effects can be. For any bad fluent f_b holding at t_1, and any good fluent f_g holding at some t_2, such that $t < t_1, t_2 \leq H$, the following holds:

$$\Gamma \vdash \begin{pmatrix} \mathbf{I}_S\big(a, t, Holds(f_b, t_1)\big) \wedge \rhd\big(Holds(f_b, t_1), Holds(f_g, t_2)\big) \\ \vee \\ \neg \rhd \big(Holds(f_b, t_1), Holds(f_g, t_2)\big) \end{pmatrix}$$

7. Proof Sketch for the $\mathcal{DTE}$

We now apply our formal definitions for primary intentions and $\mathcal{DTE}$ from above to a brief proof-sketch for the Loop Case. Drawing from Ref. 1 and the Push Case, we know that it is ethically impermissible to push someone onto the track to stop the trolley. Additionally, we know that we have an intention to save the pair of people, P_1 and P_2, on track $track_1$. Intuitively, we know that to stop the trolley in the Loop Case, we must flip the switch and have the trolley hit P_3, or in other words we intend the trolley to hit P_3. To determine the permissibility of our flipping the switch, we need to determine whether the intention of hitting P_3 is a primary intention or a secondary one. To do this, we need to only show one of $\mathbf{G}_1 - \mathbf{G}_3$ to be false, and as such we will focus on proving the negation of $\mathbf{G}_2$:

Proof. Assume the agent a primarily intends the trolley to hit P_3. Also assume that P_3 walks off the track at t_0. The trolley will then not hit P_3 at t_1 as intended. It is given that pushing P_3 at t_x is a means to having P_3 be hit at t_{x+1}. a will therefore push P_3 at t_1 so that P_3 gets hit at t_2. However, it is also given that it is impermissible to push someone and therefore not allowed. As such, a cannot push P_3 onto the track, and therefore a can not primarily intend for P_3 to be hit. □

From this, we see that our intention of P_3 being hit is a secondary one that only occurs due to the misfortune of P_3 already being on the track. As such, we are allowed to pursue the bad effect of P_3 being hit to accomplish the good effect, the pair not being hit, as our utility of the bad effects is less than the utility of the good effects.

8. Conclusion

We now quickly summarize the primary chief contributions of this work, and end by discussing promising future lines of work. In this work, we have presented a formalization of the $\mathcal{DTE}$ within a cognitive calculi. To do this, we first created an informal definition of both the test of primary intention, $\mathbf{D_1} - \mathbf{D_3}$, as well as for the $\mathcal{DTE}$, $\mathbf{C_1} - \mathbf{C_4}$. From this, we built the necessary formalizations, $\mathbf{G_1} - \mathbf{G_3}$ and $\mathbf{D_1} - \mathbf{D_4}$ of both. Finally, we present a proof sketch of how this formalization could be applied in determining the ethical permissibility of flipping the switch in the Loop Case of trolley problems.

For future work, there is an immediate next step of taking this formalization and building out the machinery necessary for use of the $\mathcal{DTE}$ in moral machines, such as described in Ref. 19. Indeed, a chief goal in creating formalizations for diverse moral doctrine is to allow machines to pick and choose which moral theories it should subscribe to for a given task, or even for usage within groups of people who differ on the grounds of race, religion, politics, etc. Having said that, it is important to note the work done in Ref. 20,21 that shows that robots are held to a different standard of humans, and are expected to do actions that would be questionable if done by a human. Indeed, in proceeding with formalization of these principles, and their subsequent usage of AI agents, will be necessary to conduct more emperical studies to see how a human views various principles as applied to an AI agent versus when applied to a human.

Acknowledgments

A grant from the Office of Naval Research to explore "moral competence in machines" (PI M. Scheutz) has provided indispensable support for the research reported herein. Crucial support also came in the form of a grant from the Air Force Office of Scientific Research to make possible "great computational intelligence" in AIs on the strength of automated reasoning (PI S. Bringsjord).

References

1. N. S. Govindarajulu and S. Bringsjord, On Automating the Doctrine of Double Effect, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, (Melbourne, Australia, 2017).
2. A. McIntyre, The Doctrine of Double Effect, in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta, 2004/2014)
3. F. Cushman, L. Young and M. Hauser, The Role of Conscious Reasoning and Intuition in Moral Judgment Testing Three Principles of Harm, *Psychological science* **17**, 1082 (2006).
4. M. Hauser, F. Cushman, L. Young, R. Kang-Xing Jin and J. Mikhail, A Dissociation Between Moral Judgments and Justifications, *Mind & Language* **22**, 1 (2007).
5. F. M. Kamm, *Intricate Ethics: Rights, Responsibilities, And Permissible Harm* (Oxford University Press, New York, New York, 2007).
6. A. McIntyre, Doing away with double effect, *Ethics* **111**, 219 (2001).
7. S. M. Liao, The loop case and kamm's doctrine of triple effect, *Philosophical Studies* **146**, p. 223–231(Jul 2008).
8. L. M. Pereira and A. Saptawijaya, *Counterfactuals in Critical Thinking with Application to Morality*, in *Model-Based Reasoning in Science and Technology*, (Springer International Publishing, 2016), p. 279–289.
9. E. T. Mueller, *Commonsense Reasoning: An Event Calculus Based Approach*, 2 edn. (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2015).
10. G. Gentzen, Untersuchungen über das logische Schlieben I, *Mathematische Zeitschrift* **39**, 176 (1935).
11. K. Arkoudas and S. Bringsjord, Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task, in *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)*, eds. T.-B. Ho and Z.-H. ZhouLecture Notes in Artificial Intelligence (LNAI)(5351) (Springer-Verlag, 2008).
12. S. Bringsjord, N. Govindarajulu, D. Thero and M. Si, Akratic Robots and the Computational Logic Thereof, in *Proceedings of ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology)*, (Chicago, IL, 2014).
13. P. Foot, The problem of abortion and the doctrine of double effect, *Oxford Review* **5**, 5 (1967).
14. J. J. Thomson, The trolley problem, *The Yale Law Journal* **94**, p. 1395(May 1985).
15. M. Hauser, F. Cushman, L. Young, R. Kang-Xing Jin and J. Mikhail, A Dissociation Between Moral Judgments and Justifications, *Mind & Language* **22**, 1 (2007).
16. M. Otsuka, Double effect, triple effect and the trolley problem: Squaring the circle in looping cases, *Utilitas* **20**, p. 92–110(Feb 2008).
17. M. E. Bratman, Intention, plans and practical reason, **100**(01 1987).
18. S. Bringsjord, A 21st-Century Ethical Hierarchy for Robots and Persons: $\mathcal{EH}$, in *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, (Lisbon, Portugal, 2017).
19. P. Bello and S. Bringsjord, On How to Build a Moral Machine, *Topoi* **32**, 251 (2013), Preprint available at the URL provided here.
20. B. Malle, M. Scheutz, T. Arnold, J. Voiklis and C. Cusimano, Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents, in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI'15*, (ACM, New York, NY, 2015) pp. 117–124.
21. B. F. Malle, S. T. Magar and M. Scheutz, AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma, in *Robotics and Well-Being*, eds. M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi and E. E. KadarIntelligent Systems, Control and Automation: Science and Engineering (Springer International Publishing, Cham, 2019) pp. 111–133.

ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.008

SIMILARITIES IN RECENT WORKS ON SAFE AND SECURE BIOLOGY AND AI RESEARCH

Pedro Henrique Oliveira dos Santos

Instituto de Informática, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500
Porto Alegre, Rio Grande do Sul 91501-970, Brazil
E-mail: pedrosans@gmail.com
http://www.inf.ufrgs.br

Dante Augusto Couto Barone

Instituto de Informática, Universidade Federal do Rio Grande do Sul
E-mail: barone@inf.ufrgs.br
http://www.inf.ufrgs.br

Given the growth rate of artificial intelligence capabilities, it's natural to pay attention to the risk involved in such unprecedented technology growth, how it affects society today and potential security threats. Such concern is not new and has been the object of scrutiny over the years due to the same threats posed by advances in biology. This article resumes recent and relevant works on both areas and lists their similarities.

Keywords: Biorisk; Artificial Intelligence; Ethic.

1. Recent Works on AI

Technologies made possible by advances in AI can greatly benefit society like by aiding disease diagnosing while giving neuroscientists a better understanding of the brain.[1] As disease diagnosing is thought as a specialized and intelligent task, the artificial mean to perform it, aided by the use of technology, is referred to as AI.[2] Because new technologies enabled by AI research is also creating new risks, like a fail in the self-driving ending in a fatality, a number of institutions are raising concern. To put in perspective, an accident involving a vehicle with autopilot hardware in March 2018[3] is being investigated for failures in the car software. To address new risks made possible by the developments in AI, a number of institutions are coming together to raise questions and create guidelines for AI systems. In February 2017 the Future of Humanity Institute reported on the current AI capabilities, how it can be used with malicious intent, and how this risk can be addressed; it produced the document The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.[2] On April 2017 UK's House of Lords Select Committee on Artificial Intelligence, aiming to lead the international community, proposed five principles for upcoming ethical AI frameworks.[4] On December 2017 IEEE published the second version of the document Ethically Aligned Design which is part of its global initiative on ethics of autonomous and intelligent systems to identify and find consensus on issues of transparency, accountability, and algorithmic bias in implementations of AI systems.[5] In this scenario, US government officials and researchers came together to better understand the possibilities of mutual support that will need to be put in place in the case of a cyber attack and reported their work in the Cyber Mutual Assistance Workshop Report.[6] As a result of the mentioned works some threats were identified, including:

- Repurpose of existing AI systems by terrorists like drones or autonomous vehicles programed deliver explosives and cause crashes

- Automating influence campaigns leveraged by IA analysis of social networks
- Denial-of-information attack leveraged by AI enabled bots capable of publishing false or distracting information next to real information
- Fake news reports with realistic fabricated video and audio
- Many jobs can disapear
- Unclear liability when AI systems malfunction or cause harm to users

Given the new risks and capabilities being made possible by AI, ethical concerns were raised in the mentioned works, including:

- Ethical design should be an integral part of the curriculum of AI system developers and users
- Well being should be prioritized in AI systems
- If an accident involving an autonomous car occur, the AS will need to be transparent to an accident investigator
- Transparency is important to understand what AI systems are doing and why
- Record-keeping of intended use and system parameters to enable investigators to find out the legally responsible for a particular AI system
- Tailored defenses for an attack will need financial backing

2. Recent Works on Biorisk

In September 2017 the Future of Life Institute published three paper to assess global catastrophic and existential biosecurity risk. In the papers, it's notable the concern with financial aspects, the benefits of threat-mitigation efforts and questions on ethics including:

- Should a dual-use research be funded?
- What should be the price for expected risks in a dual-use research?
- Should an estimated price be included as a cost in research grant proposal?

A central point in the ethical discussion on biology is the gain-of-function experiments since a subset of such experiments can be used by malicious actors.[7] To ethically balance the need for better public health by better understanding viruses, and to protect the same public from the risks associated with this research proved a point of controversy. Selgelid[8] developed/proposed in 2016 a framework for gain-of-function research decision making supported by a set of principles - including manageability of risks, justice, engagement - designed to indicate ethically acceptable or ethically problematic ou unacceptable researches. A similar initiative came from the Obama government that tasked, in 2014, the US National Science Advisory Board for Biosecurity (NSABB) to recommend on the deliberative process regarding risks associated with gain-of-function researches. In 2016 NSABB published its recommendations for evaluating gain-of-function researches supporting its values[9] with the Belmont Report,[10] the literature on public health ethics[11],[12] and the ethics analysis by Dr. Michael Selgelid.[8] The recommended ethical values include:

- Non-maleficence: research should consider and apply approaches preventing harm and mitigating potential risks.
- Beneficence: the research should have a beneficial outcome for public health
- Social justice: benefits and risks should be fairly distributed, even on a global scale if it's the case
- Accountability: actions should have a responsible actor and a justification
- Transparency: Uncertainties, controversies, and limitations should be made public and updated as the research develops

3. Finance

The Department of Homeland Security (DHS) lists financial service[13] as one of the 16 critical infrastructure sectors for USA. While the work to seek safe forms of utilizing new technologies does have a cost, in an effort exercise the security of financial systems, the Norwich University worked on a $9.9 million contract, in 2013, to develop the Distributed Environment for Decision-Making Exercises – Financial Sector (DECIDE-FS) tool, awarded by Cyber Security Division of the DHS Security Science and Technology Directorate (CMAWR).[6]

Even though big investments in safe usage of software can be found, The Malicious Use of Artificial Intelligence report does question if existing funding strategies, like to put a bounty on a vulnerability, should be extended to AI technologies, and if such bounties could be offered by third parties like government or philanthropic sources.[2]

The same financial concern can be seen in the papers on biorisk published by the Future of Life Institute. The risk of human extinction is presumably low and reducing this risks chance does have a cost. Historically, costs in public health can be high, like $13 billion on health security-related programs in 2017 projected by US federal government.[14] In the work of Millett,[15] the cost of existential risk prevention is not cost-effective when compared to basic healthcare investments, but it does show up as cost-effective when compared to the benefit of the investment. An initial estimate shows that it takes around 10 cents of a dollar to save 1 life-year. Farquhar[7] goes into details on how to price and charge, were by identifying liability in case of catastrophe and by assessing the risk of a research, research institutions can pay upfront for the risks and be incentivized to minimize its chances of happening. The fair price can be pursued by borrowing methods used by insurance companies that are already pricing risks such as cyberattack.[16]

4. Similarities

Given the work on safety and security in both biology and AI, the following similarities can be listed:

(i) Dual-use research: the concern shown by gain-of-function research funders, that the research can be used with malicious intent, can be seen in recent works of major AI researchers when questioning the risk of existing AI systems being repurposed by terrorist
(ii) Ethical framework: the need for a methodology and a set of parameters and values to reconcile ethical concerns in AI systems, being worked by the IEEE 7000 project,[17] where the goal of the NSABB report published in 2016 on gain-of-function research
(iii) Financial backing: the decision to fund a research in gain-of-function opened the same questions that can be made for dual-use research in AI
(iv) Public interest: the motivation to fund a research, even after its risk assessment, can be seen in both areas. Gain-of-function researches pursued better understanding the virus H1N1 and aided the management of the 2009 pandemic,[7] where the project 7010 by IEEE[18] elaborate well-being indicators also aiming at the public interest.

Given the similarities, it's natural to see the common response where researchers and their institutions came together to ensure the safe development of their areas. The work on safe and secure research on biology had an earlier start and is in a more mature state, already substantiating research fundings cuttings. As outlined in this article, there is a good number of recent initiatives on safe and secure AI research which points to a future that can be beneficially permeated by the technologies we are researching today. To mitigate the risks involved it's important to develop and mature the work on creating guidelines and rules for ethically aligned AI systems and the work on biology can be a reference and inspiration.

Bibliografia

1. B. Gonzales, Riascos, *How Artificial Intelligence is Supporting Neuroscience Research: A Discussion About Foundations, Methods and Applications*, tech. rep. (2017), https://link.springer.com/chapter/10.1007/978-3-319-71011-2_6.
2. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, tech. rep. (2018), https://arxiv.org/pdf/1802.07228.pdf.
3. J. Stewart, Tesla's autopilot was involved in another deadly car crash (2018), https://www.wired.com/story/tesla-autopilot-self-driving-crash-california/.
4. U. Parliament, Ai in the uk: ready, willing and able? (2017), https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm.
5. *Ethically Aligned Design A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, tech. rep., IEEE (2017).
6. *Cyber Mutual Assistance Workshop Report*, tech. rep., Carnegie Mellon University (2018).
7. O. C.-B. Sebastian F and A. Snyder-Beattie, Pricing externalities to balance public risks and benefits of research (2017), https://www.liebertpub.com/doi/pdfplus/10.1089/hs.2016.0118.
8. M. J. Selgelid, Sebastian farquhar, owen cotton-barratt, and andrew snyder-beattie (2016), https://link.springer.com/content/pdf/10.1007%2Fs11948-016-9810-1.pdf.
9. *Recommendations for the Evaluation and Oversight of Proposed Gain-offunction Research*, tech. rep., National Science Advisory Board for Biosecurity (2016), http://www.iucn-whsg.org/sites/default/files/People,%20Pathogens%20and%20Our%20Planet.pdf.
10. *The Belmont Report*, tech. rep., Department of Health, Education, and Welfare. (1979), https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf.
11. N. Kass, An ethics framework for public health. (2001), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1446875/.
12. New directions. the ethics of synthetic biology and emerging technologies. (2010), http://bioethics.gov/sites/default/files/PCSBI-Synthetic-Biology-Report-12.16.10_0.pdf.
13. Critical infrastructure sectors (2017), https://www.dhs.gov/critical-infrastructure-sectors.
14. People, pathogens and our planet, volume 2: The economics of one health. (2012).
15. P. Millett and A. Snyder-Beattie, Existential risk and cost-effective biosecurity (2017), https://www.liebertpub.com/doi/pdfplus/10.1089/hs.2017.0028.
16. M. M. Daniel Garrie, Cyber-security insurance: Navigating the landscape of a growing field (2014), https://repository.jmls.edu/cgi/viewcontent.cgi?article=1766&context=jitpl.
17. 7000 - model process for addressing ethical concerns during system design (2016), https://standards.ieee.org/develop/project/7000.html.
18. 7010 - wellbeing metrics standard for ethical artificial intelligence and autonomous systems (2017), https://standards.ieee.org/develop/project/7010.html.

ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.004

Drones and Data Protection Issues

Nicola Fabiano
Studio Legale Fabiano, Via Luigi Tosti 3,
Rome, 00179, Italy
E-mail: info@fabiano.law
www.fabiano.law

The European Regulation 2016/679 (General Data Protection Regulation - GDPR) is a revolution because it changes the perspective on this theme radically. The GDPR provides principles to protect the rights and freedoms of natural persons. Technicians, very often, ignore any legal references paying attention almost entirely to the scientific aspects. Indeed, the GDPR applies to any sector everywhere equally. Regarding drones, apart from the civil liability as a traditional area, the GDPR has a real impact on this domain both for the respect of the law provisions and for the concrete adoption of the principles, especially during the design phase. This contribution presents some of the aspects related to drones and data protection according to the GDPR.

Keywords: Drones, Data Protection, Privacy, Security.

1. The European legal framework on drones

In Europe, the legislation about drones is the REGULATION (EC) No 216/2008 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 20 February 2008 on common rules in the field of civil aviation and establishing a European Aviation Safety Agency, and repealing Council Directive 91/670/EEC, Regulation (EC) No 1592/2002 and Directive 2004/36/EC.[1]

There is a Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on common rules in the field of civil aviation and establishing a European Union Aviation Safety Agency, and repealing Regulation (EC) No 216/2008 of the European Parliament and of the Council.[2]

Generally speaking, it is possible to distinguish between manned and unmanned aircraft. Drones are classified as "unmanned aircraft" (UA).

Apart from the before mentioned European legislation, there are the domestic ones and the specific regulation issued by the International Civil Aviation Organization (ICAO) and Aviation National Authorities.

The before mentioned proposal for Regulation introduces a large number of significant innovation and especially about the definitions because it identifies the UA[a]. Moreover, the proposal describes the safety information.

The EASA[3] clarified that for drones it is possible to distinguish two main types of risks:

- air risks (collision with a manned aircraft or another UA); and
- ground risks (collision with persons or critical infrastructure)

Pinpointing the risks it is crucial to prevent the consequences of civil liability. In fact, drones can also imply civil liability in case of damages occurred to someone or something.

[a]Article 2:
1. "unmanned aircraft system (UAS)" means the unmanned aircraft (UA) and the equipment to control the UA remotely;
5. "unmanned aircraft (UA)" means any aircraft operating or designed to operate autonomously or to be piloted remotely without a pilot on board.

It is necessary to have insurance to cover all the risks that a drone can potentially cause.

2. The European Law on the processing of persona data

Apart from the aforementioned European law about drones, we have to mention the EU Regulation n. 2016/679 (General Data Protection Regulation - GDPR).[4]

In Europe, the protection of natural persons about the processing of personal data is a fundamental right. In fact, the Article 8 of the Charter of Fundamental Rights of the European Union (the 'Charter')[5] is related to the protection of natural persons about the processing of personal data[b].

Furthermore, the Charter considers also the respect for private and family life[c] as a crucial aspect of privacy.

Moreover, the Treaty on the Functioning of the European Union (TFEU)[6] considers the right to the protection of personal data[d].

In 2016 it has been published the European Regulation number 2016/679 that entered into force on 25 May 2016, but it applies from 25 May 2018.[4] According to the Article 94, this Regulation will repeal the previous European legislation on data protection (Directive 95/46/EC[7]) with effects from 25 May 2018.

The GDPR obviously mentions the Charter of Fundamental Rights of the European Union in the first Whereas[e].

The primary goal of the EU Regulation 2016/679 is to harmonise the legislation of each Member State: the GDPR will be directly applicable in each European State, avoiding possible confusion among the domestic law. The GDPR introduces numerous changes, such as the Data Protection Impact Assessment (DPIA), the Data Protection by Design and by Default (DPbDbD), the data breach notification, the Data Protection Officer (DPO), the very high administrative fines in respect of infringements of the Regulation, and so on.

Regarding the protection of personal data, apart from the before mentioned GDPR, there is also the Directive 2002/58/EC[8] concerning the processing of personal data and the protection of privacy in the electronic communications. In fact, according to the Article 95 of the GDPR, there is a relationship with this Directive[f].

The Directive 2002/58/CE has the aim to "*to ensure an equivalent level of protection of fundamental rights and freedoms, and in particular the right to privacy, with respect to the processing of personal data in the electronic communication sector and to ensure the free movement of such data and of electronic communication equipment and services in the Community*"[g].

In this legal panorama, it is clear that technology and law are not at the same level

[b] Article 8 - Protection of personal data. 1. Everyone has the right to the protection of personal data concerning him or her. 2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified. 3. Compliance with these rules shall be subject to control by an independent authority.

[c] Article 7 - Respect for private and family life. Everyone has the right to respect for his or her private and family life, home and communications.

[d] Article 16(1) says: "Everyone has the right to the protection of personal data concerning them".

[e] The protection of natural persons in relation to the processing of personal data is a fundamental right. Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) provide that everyone has the right to the protection of personal data concerning him or her.

[f] The Article 95 says: "This Regulation shall not impose additional obligations on natural or legal persons in relation to processing in connection with the provision of publicly available electronic communications services in public communication networks in the Union in relation to matters for which they are subject to specific obligations with the same objective set out in Directive 2002/58/EC".

[g] Article 1

because the first one (technology) is always ahead than the second one (law). The actions on the part of the legislator always followed the technological solutions, and so the law rules have to be able to consider the technology evolution.

It is crucial to analyse the GDPR to be ready and comply with the new data protection Regulation. In fact, the General Data Protection Regulation (GDPR) represents an innovative data protection law framework, because of several purposes on which is based and strictly related to the technical solutions.

2.1. *Data Protection and Privacy*

Very often people talk about data protection, but using the term "privacy" as a synonym, confusing the real meaning indeed. "Privacy" and "Data Protection" are not the same because, apart from the definition, they are different concepts. Both are fundamental rights in Europe, but there are differences between them.

On the one hand privacy is related to the personal life; on the other hand, data protection concerns the protection of natural personal about the processing of personal data.

It is not possible to address data protection and privacy issues adopting only technical solutions without any legal reference. Apart from the highly technical measure, hence, we cannot dismiss the law obligations, where they are applicable, like in Europe, according to the GDPR.[4] In fact, an approach considering the legal framework, confirms the equation according to security is not equal to privacy. A system could be very secure but not compliance with the data protection law. On the contrary, a system could be compliance with the data protection law and, hence, very secure (obviously only by the adoption of security measures).

3. Drones and Data Protection

One of the main legal issues is related to the personal data protection law (GDPR). First of all, it is essential to highlight the **territorial scope** according to the article 3 of the GDPR, because the EU Regulation 2016/679 applies in all over the world regardless of whether the processing takes place in the Union or not. In fact, just according to the article as mentioned earlier, 3, paragraph 1, "*This Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of* ***whether the processing takes place in the Union or not***". Moreover, the paragraph 2 of the same article 3, states "*This Regulation applies to the processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union, where the processing activities are related to:*

(a) *the offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the Union; or*
(b) *the monitoring of their behaviour as far as their behaviour takes place within the Union*".

Regarding this topic, the European Data Protection Board (EDPB) carried out the "Guidelines 3/2018 on the territorial scope of the GDPR - Adopted on 16 November 2018".[9]

Apart from the territorial scope, we must distinguish the use of drones from designing of them. In case of use of drones, the user has to pay attention to the data protection and privacy laws, according to which people have to respect data subject's rights and personal life.

Regarding the use of drones, for example, very often people install cameras to acquire photo or video, but according to the law, it is allowed to acquire images or video only in a public place. In fact, it is not permitted to obtain pictures of a natural person in a private

area without his/her consent, and this can be the case of unlawful processing of personal data.

Regarding the design of drones, we have to apply other rules laid down in the GDPR and especially the principle "data protection by design" according to the article 25 of the GDPR, where any infringement of this provision shall be subject to administrative fines up to 10.000.000 EUR, or in the case of an undertaking, up to 2% of the total worldwide annual turnover of the preceding financial year[h]. The principle laid down in the Article 25, paragraph 1, of the GDPR is the data protection by design, and it means that during the design phase the controller[i] "*shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects*". This principle means that the controller is responsible for, and can demonstrate compliance with, the principles contained in the Article 5, paragraph 1, of the GDPR.

Moreover, the user of drones (controller or processor) has to respect the security measures according to the GDPR, by implementing "*appropriate technical and organisational measures to ensure a level of security appropriate to the risk*[j]".

One of the main risks is the unlawful collection of personal data.[10]

Ethics entails a consciousness about the high values belonging to a natural person. People have to respect human dignity and Ethics both using or designing a drone. It is not simple to define Ethics, but we believe that it is essential to raise awareness. The risk is that natural person becomes pure data, debasing and losing so the typical aspects belonging to a human. Ethics is the correct path to preserve the ontological nature of human.

What is ethics?

There are no easy answers because we have several definitions. We want to refer to a thinking way helpful to distinguish, generally speaking, what is wrong from what is right, finding the right key to conferring a natural person the exact value belonging to him or her.

The GDPR doesn't lay down any specific rules on Ethics. Nevertheless, we think that it is possible to start applying the GDPR principles thinking ethical: it is a matter of approach even without any norm.

Apart from the laws, in the processing of personal data, each controller should consider ethics anyway also even does not exist any obligation provided by the law.

A consciousness of ethics entails the comply with the law (data protection law - GDPR), but it could not true the contrary (respecting the law does not mean to have always consciousness of ethics).

4. Conclusion

Drones can imply risks for natural persons regarding the protection of personal data.

The protection of personal data is entirely relevant not only regarding the use of drones and the consequences of civil liability but also during the design phase. In fact, designing drones technicians have to pay attention to the data protection by design and by default

[h] Article 83, paragraph 4 letter (a) of the GDPR

[i] The Article 4, paragraph 1, of the GDPR defines the controller as "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law"

[j] Article 32 of the GDPR

principles according to the GDPR. However, to develop drones, it could be used Artificial Intelligence and Machine Learning. Nowadays it is possible to design drones as a robot, and in this way, it is crucial to consider the relationship between ethics and robots fully.

Ethics and robots cannot dismiss from the data protection law and the protection of the risk of varying likelihood and severity for the rights and freedoms of natural persons.

References

1. REGULATION (EC) No 216/2008 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 20 february 2008 on common rules in the field of civil aviation and establishing a european aviation safety agency, and repealing council directive 91/670/eec, regulation (ec) no 1592/2002 and directive 2004/36/ec (2008), http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02008R0216-20130129&from=EN.
2. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on common rules in the field of civil aviation and establishing a european union aviation safety agency, and repealing regulation (ec) no 216/2008 of the european parliament and of the council (2015), http://eur-lex.europa.eu/resource.html?uri=cellar:da8dfec1-9ce9-11e5-8781-01aa75ed71a1.0001.02/DOC_1&format=PDF.
3. E. A. S. Agency, Opinion no 01/2018 - introduction of a regulatory framework for the operation of unmanned aircraft systems in the 'open' and 'specific' categories (2018), https://www.easa.europa.eu/sites/default/files/dfu/Opinion%20No%2001-2018.pdf.
4. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (2016), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN.
5. Charter of fundamental rights of the european union (2012), http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT&from=EN.
6. The treaty on the functioning of the european union (2012), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012E/TXT&from=EN.
7. DIRECTIVE 95/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (1995), http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&from=EN.
8. DIRECTIVE 2002/58/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 12 july 2002 concerning the processing of personal data and the protection of privacy in the electronic communications (2002), http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32002L0058&from=en.
9. Guidelines 3/2018 on the territorial scope of the gdpr (2018), https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_3_2018_territorial_scope_en.pdf.
10. Office of the Information Commissioner Queensland, GUIDELINE Drones and the Privacy Principles (2018), https://www.oic.qld.gov.au/guidelines/for-government/guidelines-privacy-principles/applying-the-privacy-principles/drones-and-the-privacy-principles.

ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.023

REVISITING THE CONCEPT OF [WORK] IN THE AGE OF AUTONOMOUS MACHINES

MARIA ISABEL ALDINHAS FERREIRA

Centre of Philosophy of the University of Lisbon- Language, Mind and Cognition Group
Faculty of Arts and Humanities. University of Lisbon
and
Institute for Systems and Robotics, Instituto Superior Técnico. University of Lisbon

isabelferreira@letras.ulisboa.pt

Abstract:

At the dawn of the overspread deployment of autonomous systems in most if not in every domain of life, the present paper revisits the concept of [work] claiming that the capacity for work is an essential human attribute. As it happens with all the attributes that define distinct life forms, this attribute has evolved all along human developmental history, as a consequence of their adapting and responding to distinct environmental physical conditions, to distinct modes of production and consequently to differentiated social and cultural contexts. Being an essential specific attribute of this social species and not the result of a temporary condition or state, [work], due to its generative power, is responsible for guaranteeing individual and collective subsistence throughout times, the evolution of distinct socio-economic frameworks, the emergence of rudimentary and progressively more sophisticated tools, the creation of culture and art forms and in so doing has been playing a determining role on the very evolutionary and developmental processes it emerges from.

Contemporary societies are becoming progressively more and more hybrid environments. This means environments where the physical is permeated by the digital, where human interaction is mediated by advanced forms of communication, where non-embodied and very soon also embodied forms of artificial intelligence coexist with natural intelligence[1] where ultimately [work] in its intrinsic humaness is being replaced by task performing by artificial autonomous systems.

This present context and its predictable development in the near future[2] demand the emergence of a deep awareness on the part of policy makers and from society in general so that technology remains a tool for enhancing [work], respecting its fundamental twofold dimension as: (i) a human generative endowment for the creation and transformation of reality (ii) a means for every human being to effectively be a part of this reality in all their dignity.

Keywords: Tools, Work, Technological Development, Hybrid World, Human Dignity

1. Introduction

In a significant number of species, the evolutionary and developmental process has proceeded according to three interconnected axes: (i) interaction ability, (ii) task performance ability, (iii)

[1] We are referring not just to the form of intelligence that characterises human cognition, but the ones that are inherent to all the other life forms

[2] According to the International Federation of Robotics, more than 1.7 million industrial robots will have been installed in factories, worldwide, by 2020 and about 400 million workers around the globe will be displaced by 2030
https://ifr.org/ifr-press-releases/news/ifr-forecast-1.7-million-new-robots-to-transform-the-worlds-factories-by-20

tool making ability[3]. The abilities represented by these three axes are made possible by a set of innate endowments that, though exhibiting a degree of variability among species in what relates their level of sophistication and complexity, represent a continuum that is horizontal to all of them. This way the capacity for communication, that the interaction ability subsumes, attains its highest degree of sophistication in human language, the same happening with the capacity for distributed task performing and the tool making capacity which are endogenous[4] and define a continuum of progressive complexity throughout different species[5]. In what concerns human beings, these abilities are also the result of a learning process that takes place in society, that goes on throughout the individual's lifetime and that relies substantially on the accumulated experience of precedent generations and on the prevalent economic, social and cultural models. One of the substantial differences we can immediately identify when contrasting tool making in humans and in other species is the fact that while human tools have been evolving exponentially - with some tools exhibiting nowadays a considerable degree of potential autonomy that will require less or none human intervention - tool making among other species has remained rudimentary.

Tool making is inherently associated to the biological and the social and cultural evolution and development of human kind. Throughout the ages, human beings have modified or updated the inventions of precedent generations or those of other communities of tool makers and have also created new ones in order to achieve certain goals. From distinct modes of production specific scientific and technological innovations have emerged determining distinct working settings, distinct working tools.

Marx refers to tool use as an extension of the laboring body[6] and views technologies as extensions of the human will domination over nature:

"Nature builds no machines, no locomotives, railways, electric telegraphs […]These are products of human industry; natural material transformed into organs of the human will over nature…they are organs of the human brain created by the human hand" Marx 1993:706

Tool making and its natural evolution is inherently associated to the huge transformative and generative power resulting from the endogenous human working capacity.

2. Artifacts for Work: Tools

Human beings have been producing millions of artifacts, characteristic of particular civilizational frameworks and stages of development. As it happens with all the entities that populate the human world, each of these artifacts is value-laden, i.e., they have got an identity realized by a semantic value that defines and determines this same identity in the context of their use in particular social and cultural settings. But independently of this setting or context of use, there is a trait common to all artifacts. That feature is {function} which identifies the purpose an object fits in. It is the definition of this feature that allows us to distinguish [table] from [chair] or [glass] from [bottle]. And though the two first ones are comprehended in the

[3] The interaction ability, in fact, subsumes either (ii) or (iii). We make the distinction for purely analytical purposes

[4] Greenfield P. M. (1991). Language, tools, and brain: the development and evolution of hierarchically organized sequential behavior. Behav. Brain Sci. 14, 531–595

[5] Although tool use was for long been assumed to be a uniquely human trait, there is now much evidence that other species such as mammals, namely primates, birds, cephalopods also use more or less rudimentary tools. Cf. Shumaker, R.W., Walkup, K.R. and Beck, B.B., (2011) and Beck, B.B. (2013)

[6] Grigenti, Fabio (2016)

broader category of [furniture] and the later in the broader categories of [container] and probably of [glassware] no one will, certainly, have any difficulty, under normal conditions, to distinguish the first from the others.

Tools are a particular subset of artifacts[7]. They share, with the broader category they belong to, the feature {function}, i.e. they are suited to a particular purpose, but their semantic specificity is realized by another fundamental feature. When looking at the definition of the concept [tool] in a language dictionary[8] we read:

i. A **tool**[9] is any instrument or piece of equipment that you hold in your hands **in order to help you to do a particular kind of work.** e.g., "workers downed tools in what soon became a general strike".

ii. A **tool** is also any object, skill, idea etc, **that you use in your work** or that you need for a particular purpose.

By analyzing these definitions we realize that the concept of [tool] is primarily associated to a working scenario and to the production/creation of a particular entity. This means that inherent to the semantics grounding the concept of [tool] is the trait **{cause an object, an event or a state to come into being, through physical and/or mental activity}** which is the essence of the concept of [work], weather the nature of this work is tangible or not. This fact can be easily foreseen when we think not only of a shoemaker handling their tools to create or repair a pair of shoes; a dressmaker making a dress, a carpenter making a chair, but also of a factory worker interacting with a machine to produce a particular piece, the farmer that drives a tractor to plough the field, the researcher that sits at the computer using a text processor to write a paper…

Tools can be seen as body extensions[10] not only in the sense that they provide a means for accomplishing an action that the bare corporeal architecture was not able to perform by itself, but essentially because handling and/or operating any kind of tool always requires the adoption of specific protocols associated to the definition of new neural pathways that frequently comprehend the triggering out of specific motor programmes[11] enabling particular postures or body movement. Tool handling is consequently responsible for the definition of neural pathways that will allow particular patterns of behavior to become typical and routinary, being, this way, instantly triggered out by specific contexts of use without depending on a reflexive attitude[12].

---

[7] The definition of the concept of [tool] has been subject to different versions by researchers studying animal behavior.Hauser (2000) defines [tool] as an object that has been modified to fit a purpose or an inanimate object that one uses or modifies in some way to cause a change in the environment, thereby facilitating one's achievement of a target goal.

[8] Collins Cobuild English Language Dictionary. Collins Publishers. University of Birmingham 1988 (1st edition)

[9] Emphases mine

[10] It is particularly interesting how some technological artifacts or technological tools are sometimes presented as extensions of the physical body. We recall on this purpose a small video opening a laptop computer produced by Texas Instruments in the early 90's - Texas Extensa- that stated: " Texas Extensa, an Extension of yourself"

[11] According to Young, R. (2003) the two fundamental human handgrips, first identified by J. R. Napier, and named 'precision grip' and 'power grip', represent a *throwing grip* and a *clubbing grip*, thereby providing an evolutionary explanation for the two unique grips, and the extensive anatomical remodelling of the hand that made them possible.

[12] Cf Ferreira (2014)

This prosthetic nature of the tool is also addressed by Heidegger (1962) that refers how tools are taken into ways into which human beings enroll and project themselves into work practices as they "withdraw" and become "ready-to-hand"

Perhaps because of this nearly physiological extension, this quasi-symbiotic process, between a human being and a specific instrumental artifact, through which a specific entity is produced and comes into being, there is frequently a link of affective attachment that unites workers to their tools and to the produced works. This frequent affective attachment reflects itself in the care often revealed by workers in the maintenance and keeping of their tools[13], in the way artisans have always carved out or just signed their names on the created object or in the sense of achievement and even pride manifested by those that have contributed to the coming into being of important realizations. This feeling was recently evident, for instance, in a newspaper interview to the workers that participated in the construction of the 25th of April Bridge (former Salazar Bridge) in Portugal, on the occasion of its 50 anniversary[14]. According to António Rosa, one of these workers, the construction of this bridge, the biggest in Europe at that time, was a real challenge to everyone and it came out becoming a kind of second home to those deeply committed to their construction. With more than 40 years dedicated first to its building and then to its maintenance, this worker confessed that he still kept some of the tools he used then, namely a brush.

3. Work as a Human Endowment

Different ideological perspectives, distinct epistemological frameworks[15] have converged on recognizing the uniqueness of [work] as a human endowment and its essential character in the definition of what to be human means.

To Marx (1968), [work] is the unique means through which human beings objectify their existence and come into being. It is this essential objectivation that sustains human condition and is the essence of humanity.

In fact when we reflect on the process through which individual identity is shaped we realize that it develops according to a succession of social circles[16], starting in infancy, with the very small circle of close family members, and progressively broadens to others circles (friends and acquaintances, school/academic circle, working/professional circle…) that partially overlapping define the individual's essence. In this process of identity formation, which Ferreira (2007, 2011) has compared to the process of formation of a pearl, the working/professional circle is fundamental for the definition of the role(s) the individuals will play in the social tissue and the way they will participate and act on it. What the individual is depends in fact on what s/he does, i.e., depends on the nature of their contribution, of their work embedded in the specific social context they belong to at a given historical time.

Another fundamental perspective on the essential character of this endowment in the definition of humaness is the Encyclica Laborem Exercens (14th September 1981). This encyclical, written by Pope John Paul II, is part of the larger body of Catholic social teaching tracing its origin back to Pope Leo XIII's 1891 encyclical Rerum Novarum. The Encyclica Laborem Exercens highlights the fact that the capacity for work is an essential human feature

[13] We recall on this purpose the particular attachment an hairdresser revealed towards her set of high specialized scissors, which she had acquired when becoming a professional and uses in her daily practice

[14] https://www.sabado.pt/portugal/detalhe/conhece-o-dono-da-ponte-25-de-abril

[15] Cf on this purpose Marx (1968) and John Paul II (1981)

[16] Cf Ferreira, (2007) (2011)

that cannot be comparable by its intrinsic characteristics to the performing of certain tasks by other species in order to subsist.

"Work is one of the characteristics that distinguishes man from the rest of creatures, whose activity for sustaining their lives cannot be called work [….] it bears a particular mark of man and of humanity, the mark of a person operating within a community of persons. "(ibidem:1)

As the encyclical points out [work] is universal in the sense that it embraces "all human beings, every generation, every phase of economic and cultural development" and it is simultaneously a process that takes place within each human being, a personal narrative acknowledged by the conscious subject. Consequently it develops along two fundamental inseparable and complementary dimensions:

i. An objective dimension

ii. A subjective dimension

Its objective dimension relates to its generative and transformative power through which human beings act on the surrounding environment- "dominating nature"[17], "subduing the earth"[18]- and by so doing creating with the effort of their bodies and the intelligence of their intellects the necessary conditions for their "being" throughout the dynamics of an existential historical time.

"[….]there thus emerges the meaning of *work in an objective sense,* which finds expression in the various epochs of culture and civilization".(ibidem, 2)

This objective dimension is the tangible or non-tangible existential imprint registered not only by each society but by each of its individual members, from the most notorious to the most anonymous, since individual and collective existence and progress depend on the coordinated action and work of each and all in the different domains of human life.

John Paul II points out that [work] has an ethical value of its own, which clearly and directly remains linked to the fact that the one who carries it out is a person, a conscious and free subject.

"Working at any workbench, whether a relatively primitive or an ultramodern one, a man can easily see that through his work he enters into two inheritances: the inheritance of what is given to the whole of humanity in the resources of nature, and the inheritance of what others have already developed on the basis of those resources, primarily by developing technology, that is to say, by producing a whole collection of increasingly perfect instruments for work"(ibidem:6)

On the other hand, the subjective dimension relates to the consciousness every worker must acquire of their personal narrative and of the importance of their individual role, their contribution in a collective process to which all individual efforts converge. It is in its inherent humanity that resides the dignity of [work]:

"through work man not only transforms nature, adapting it to his own needs, but he also achieves fulfilment as a human being and indeed, in a sense, becomes "more a human being". (ibidem:9)

4. When Tools Become Autonomous Machines: The Ontological Shift

[17] Marx (1968)

[18] Laborem Exercens 1981

Technological development has resulted and results from the [work] of many thousands, throughout multiple generations, from their creativity and accumulated experience, aiming at producing the necessary conditions to liberate individuals from the toil frequently associated to hard work, promoting individual well-being and society's development, improving life conditions, eradicating poverty and disease, assuring defense against eventual threats. The huge technological development brought by the digital revolution and ICT technologies with the progressive introduction of different forms of artificial intelligence[19] in the means of production and in society in general - the 4IR - will cause an impact that is even more impressive than that brought by the first industrial revolution. Contemporary societies are in fact becoming progressively more and more hybrid environments. This means environments where the physical is permeated by the digital, where human interaction is mediated by advanced forms of communication, where non-embodied and very soon also embodied forms of artificial intelligence coexist with natural intelligence where ultimately [work], in multiple contexts and domains, is being replaced by task performing by autonomous systems.

Laying aside the evident differences inherent to the distinct stages of development that characterize the momentum of the present and those of the past technological revolutions, perhaps the most important feature brought about by the present one is that of the ***ontological shift*** of the concept of [tool]. In fact, so far, either hand tools or machine tools were manipulated or operated by human beings, depending on human skill and on their will. However [tools] are becoming progressively more and more independent from human control. By introducing forms of artificial intelligence in the means of production , by endowing machines with a form of intelligence that assigns them the capacity to operate and perform tasks independently of any form of human supervision, technology is in fact producing not tools but entities to which can easily be assigned the status of workers.

Martins,M. (2011: 18) refers to this as "a technological mutation, that ceases to be instrumental and conceived as an extension of the human arm but merges with human being, producing the very arm and threatening to produce the whole being"[20].

More than the predicted huge unemployment[21] - that in our opinion can be reverted or at least minimized by implementing the adequate social and political measures necessary to anticipate its negative impacts- e.g., prequalifying and training workers in order to their resetting- it is the potential expropriation of the generative and transformative power from human "hands/minds" that can become an existential problem.

---

[19] "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.-
AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications".in

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Artificial Intelligence for Europe {SWD(2018) 137 final}

[20] My translation

[21] Andy Haldane, chief economist at the Bank of England. predicted, in 2015, that 15 million jobs in the UK, roughly half of all jobs, were under threat from automation. He pointed out that the first industrial revolution had occurred in the middle of the 18th century and the second in the latter half of the 19th century.The third industrial evolution – the era of information technology – appeared to have resulted in an intensification of trends seen in the first two: "a hollowing-out of employment, a widening distribution of wages and a fall in labour's income share".

https://www.theguardian.com/business/2015/nov/12/robots-threaten-low-paid-jobs-says-bank-of-england-chief-economist

Hal Varian, chief economist at Google, predicted the future in the following terms: " The future is simply what rich people have today. The rich have chauffeurs. In the future, we will have driverless cars that chauffeur us all around. The rich have private bankers. In the future, we will all have robo-bankers […] One thing that we imagine that the rich have today are lives of leisure. So will our future be one in which we too have lives of leisure, and the machines are taking the sweat? We will be able to spend our time on more important things than simply feeding and housing ourselves?"[22]

These words come, in a way, nearly in line with the prediction made John Maynard Kaynes in *Economic Possibilities for our Grandchildren* (1930: I) [23]:

"My purpose in this essay [….] is not to examine the present or the near future, but to disembarrass myself of short views and take wings into the future. What can we reasonably expect the level of our economic life to be a hundred years hence? What are the economic possibilities for our grandchildren?

[…] We are being afflicted with a new disease of which some readers may not yet have heard the name, but of which they will hear a great deal in the years to come--namely, technological unemployment. This means unemployment due to our discovery of means of economizing the use of labour outrunning the pace at which we can find new uses for labour. [….]But this is only a temporary phase of maladjustment. All this means in the long run that mankind is solving its economic problem.

I would predict that the standard of life in progressive countries one hundred years hence will be between four and eight times as high as it is to-day [….] Yet there is no country and no people, I think, who can look forward to the age of leisure and of abundance without a dread. For we have been trained too long to strive and not to enjoy. […] For many ages to come the old Adam will be so strong in us that everybody will need to do some work if he is to be contented […]Three-hour shifts or a fifteen-hour week may put off the problem for a great while. For three hours a day is quite enough to satisfy the old Adam in most of us!"

Kaynes was correct in at least three of his main points:

- Unemployment due to the discovery of means of economizing the use of labour outrunning the pace at which one can find new uses for labour.
- That US economy has nearly grown 8 times the standard living in 1930.
- The need to balance a life with sufficient leisure time with the intrinsic need of the human being to produce and create, ironically referred as the "old Adam in most of us"

In fact it can easily be forseen how a non-productive existence, where human beings are deprived of an active contribution by intelligent machines, would lead to human psychological and physical degeneracy in a life devoid of purpose and dignity.

Reflecting on the complex equilibrium demanded by the relationship between human beings and machines, Gehlen(1940),(1957) refers to this relationship as a form of human empowerment. In his opinion machines do not limit human faculties as they are modeled on them but take them to a higher level. The fact that they can exalt human capacity to take physical action on our environment, even to the point of enabling new and unnatural functions (such as human flight), goes to show that machines are capable of forming part of a man-machine assembly for the purpose of going beyond boundaries previously believed impossible to overcome.

[22] https://www.ft.com/content/4329a987-9256-3059-b36f-1aba9338b800

[23]

Also referring metaphorically to the nature of this equilibrium, namely in what concerns present technology, Collins Sebastian (2018) refers that:

" Ironman isn't ironman without the suit, but the suit has no power without the man. That is the future of robotic development, people and robots, working together hand-in-hand to accomplish more than we ever thought would be possible"

On his hand, Paul Jones, professor of information science at the University of North Carolina, Chapel Hill, claims that future artificial intelligence (AI) will do well at enhancing human well-being:

"Humans need tools. Humans need and want augmentation. And as the saying goes 'First we make our tools, then our tools form us.' Since the first protohuman, this has been true".

5. Conclusions

In What New Jobs do We See Ahead?(2016:37), Elizabeth Curmi refers that in the past, automation and technological progress have not made human labour obsolete and society managed to adapt by creating new jobs to compensate for the loss of labour. As she points out, "in 1900, 41% of the US workforce was employed in agriculture; however by the year 2000, that share had fallen to 2%, mostly due to mechanisation of the sector. In the developed world, industrialisation moved people into factories and then moved them out again into services. Throughout these changes the number of jobs has always increased. US employment increased from 1950 to 2014, and the unemployment rate in 2015 (5%) is returning back to its average after the financial recession of 2008/2009"

Though today's newer technology sectors have not provided the same opportunities, particularly for less educated workers, as the industries that preceded them we cannot but agree that "technology eliminates jobs, not work"[24]

As Curmi points out, "what is clear is that technology has already changed the way we work and will continue to do so […].At all skill levels, most jobs in demand will be characterised by non-routine tasks which are not easily replaced by technology or organisational change.

We can conclude by saying that in a hybrid world, [work] will also have a hybrid nature resulting from the use of intelligent tools functioning in conjunction with natural intelligence toward the achievement of human goals. A hybrid reality where human beings will hopefully not aim to subdue or dominate Nature or their peers, but aim to be One with Nature and the rest of Humanity.

References:

Beck, B., (1980). *Animal Tool Behaviour: The Use and Manufacture of Tools by Animals* Garland STPM Pub.

Curmi, E What New Jobs do We See Ahead?in Carl Benedikt Frey, Michael Osborne and Craig Holmes *Technology at Work, v2.0 The Future Is Not What It Used to Be*, 2016. Oxford Martin School. Citi GPS: Global Perspectives & Solutions. available at: https://www.oxfordmartin.ox.ac.uk/downloads/reports/Citi_GPS_Technology_Work_2.pdf

Communication From The Commission To The European Parliament. The European Council, The European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe {SWD(2018) 137 final}

[24] Why are there still so many jobs? The History and Future of Workplace Automation, Journal of Economic Perspectives, Vol. 29, No. 3 pp 3-30.

Ferreira, M.I.A. (2007, 2011) On Meaning: Individuation and Identity. Cambridge Publishers

Gehlen, Arnold. 1940. Der Mensch seine Natur und seiene Stellung in der Welt . Berlin: Junker und Dünnhaupt.

Gehlen, Arnold. 1957. Die seele im technischen Zeitalter . Hamburg: Rowohlt Taschenbuch Verlag.

Jünger, Friedrich Georg. 1956. Die Perfektion der Technik . Frankfurt a. M: Klostermann

Grigenti, Fabio (2016) Marx, Karl- From Hand Tool to Machine Tool, in Existence and Machine. The German Philosophy in the Age of Machines (1870-1960). Springer https://www.springer.com/cda/.../9783319453651-c2.pdf..

DOI 10.1007/978-3-319-45366-8_2.

Hauser, M. (2000) The Evolution of Communication. MIT Press

Heidegger, M., (1962) Being and Time. Translated from Sein und Zeit 7th edition, Max Niemeyer Verlag. Blackwell Publishing Limited

Keynes, J. M. , (1930) Economic Possibilities for our Grandchildren http://www.econ.yale.edu/smith/econ116a/keynes1.pdf

Martins, M., L., (2011) Crise no Castelo da Cultura: Das Estrelas para os Écrans. https://repositorium.sdum.uminho.pt/bitstream/1822/29167/1/CriseCastelodaCultura.pdf

Marzke M. W. (2013) Tool making, hand morphology and fossil hominins . Philosophical Transactions of the Royal Society B. Biological Sciences. Royal Society Publishing Published 7 October 2013.DOI: 10.1098/rstb.2012.0414

http://rstb.royalsocietypublishing.org/content/368/1630/20120414#ref-list-1

Marx, Carl. 1867. Das Kapital https://www.marxists.org/archive/marx/works/download/pdf/Capital-Volume-I.pdf

Sebastian, C. (2018) Robotics and Millenials available at https://usblog.softbankrobotics.com/robots-and-millennials-joining-forces-to-change-the-future-of-work (August 2018)

Shumaker, R.W., Walkup, K.R. and Beck, B.B., (2011). Jump up to:a b c in *Animal Tool Behavior: The Use and Manufacture of Tools by Animals* Johns Hopkins University Press, Baltimore

Sikka, Sonia (2018) Heidegger Moral and Politics, questioning the Shepherd of Being.Cambridge University Press

Stout D. (2011) Stone toolmaking and the evolution of human culture and cognition. Philos Trans R Soc Lond B Biol Sci. 2011 Apr 12; 366(1567): 1050–1059.

doi: 10.1098/rstb.2010.0369. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049103/

Tinbergen, Niko (1953). The Study of Instinct. Oxford, Clarendon Press.

Young, R. W. (2003) Evolution of the human hand: the role of throwing and clubbing. J Anat. 2003 Jan; 202(1): 165–174.

doi: 10.1046/j.1469-7580.2003.00144.x https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1571064/

ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.015

AI CONCEPTUAL RISK ANALYSIS MATRIX (CRAM^{SM})

MARTIN CIUPA[1] AND KEITH ABNEY[2]
[1]*CTO calvIO Inc., Webster, New York USA*
mciupa@calvIOinc.com
[2]*Senior Lecturer, Cal Poly-SLO, California USA*
kabney@calpoly.edu

Abstract: AI advances represent a great technological opportunity, but also possible perils. This paper undertakes an ethical and systematic evaluation of those risks in a pragmatic analytical form of questions for designers/implementors of AI-Based systems. We structure this dialog as Conceptual Risk Analysis Matrix (CRAM^{SM}). We look at a topical case example in an actual industrial setting and apply CRAM^{SM}. Conclusions to its efficacy are drawn.

Key Words: AI Risk, Risk Analysis, Dialog Systems.

1. Introduction

A common worry about AI is that it poses an unacceptable risk to humanity (or individual humans) in some way. An extensive literature has begun to emerge about various aspects of Artificial Intelligence (AI) risk, much of it focused on existential risk from Artificial Generic Intelligence (AGI). But AI poses other risks, from how driverless cars solve the 'trolley problem', to whether autonomous military robots attack only legitimate targets, to trust in the safety of AI/Robotics in industrial and commercial settings. More generally, the discussion of risks from AI has paid insufficient attention to the nature of risk itself, as well as how decisions about the acceptability of the risks of AI compare to worries about convergent technologies. For example, in military robotics serious concern exists over a possible lack of "meaningful human control" [1]. Missing is a similar concern for autonomous AI-controlled cyberattacks that would lack the very same control [2]. The Vice Chairman of the Joint Chiefs of Staff understands, saying, "In the [Defense] Department, we build machines and we test them until they break. You can't do that with an artificial intelligence, deep learning piece of software. We're going to have to figure out how to get the software to tell us what it's learned" [3]. Such issues apply well beyond the military, and demand an analysis of AI risk that also applies to civilian contexts and to risks that do not rise to the level of human extinction.

So, how best to understand the risks of AI, judge them (un)acceptable, and then apply our insights on risk to determine what policies to pursue?

2. Defining risk, and how to think about it

So, AI poses many different types of risk – but what exactly is risk? Andrew Maynard [4] suggests that we start with the idea of "value." If innovation is defined as creating value that someone is willing to pay for, then he suggests risk as a *threat to value*, and not just in the ways value is usually thought of when assessing risk, such as health, the environment or financial gain/loss. The possible loss of well-being, environmental sustainability, deeply held beliefs, or even a sense of cultural or personal identity should also count. Risk's opposite, safety, should be seen as relative, not absolute: safety in all respects is never 100% guaranteed, so as safety is best understood as relative freedom from a threat of harm, so risk is a relative exposure to such a threat.

Extending a schema based on previous work [5], the major factors in determining 'acceptable risk' in AI will include (but are not limited to):

2.1 Acceptable-Risk Factor: Consent

Consent: Is the risk voluntarily endured, or not? For instance, secondhand smoke is generally more objectionable than firsthand, because the passive smoker did not consent to the risk, even if the objective risk is smaller. Will those who are at risk from AI reasonably give consent? When would it be appropriate to deploy or use AI without the meaningful consent of those affected? Would non-voluntariness (in which the affected party is unaware of the risk/ cannot consent) be morally different from involuntariness (in which the affected party is aware of the risk and does not consent)? [6]

2.2 Acceptable-Risk Factor: Informed Consent

Even if AIs only have a 'slave morality' in which they always follow orders [7], and citizens consent to their use (through, say, political means), that still leaves unanswered whether the risk (of malfunction, unintended consequences, or other error) to *unintended* parties is morally permissible. After all, even if widespread consent is in some sense possible, it is completely unrealistic to believe that all humans affected by AI could give *informed* consent to their use. So, does the morality of consent require adequate knowledge of what is being consented to?

Informed consent: Are those who undergo the risk voluntarily fully aware of the true nature of the risk? Or would such knowledge undermine their efficacy in fulfilling their (risky) roles? Or are there other reasons for preferring ignorance? Thus, will all those at risk from AI know that they are at risk? If not, do those who know have an obligation to inform others of the risks? What about foreseeable but unknown risks—how should they (the 'known unknowns') be handled? Could informing people that they are at risk ever be unethical, even akin to terrorism?

2.3 Acceptable-Risk Factor: The Affected Population

Even if consent or informed consent do not appear to be morally required with respect to some AI, we may continue to focus on the affected population as another factor in determining acceptable risk:

Affected population: Who is at risk—is it merely groups that are particularly susceptible or innocent, or those who broadly understand that their role is risky, even if they do not know the particulars of the risk? For example, in military operations civilians and other noncombatants are usually seen as not morally required to endure the same sorts of risks as military personnel, even (or especially) when the risk is involuntary or non-voluntary.

2.4 Acceptable-Risk Factor: Step risk versus State risk

A state risk is the risk of being in a certain state, and the total amount of risk to the system is a direct function of the time spent in the state. Thus, state risk is time-dependent; total risk depends (usually linearly) on the time spent in the state. So, for us living on the surface of the Earth, the risk of death by asteroid strike is a state risk (it increases the longer we're here).

Step risk, on the other hand, is a discrete risk of taking the next step in some series or undergoing some transition; once the transition is complete, the risk vanishes. In general, step risk is not time-dependent, so the amount of time spent on step matters little (or not at all). [8] Crossing a minefield is usually a step risk – the risk is the same whether you cross it in 1 minute or 10 minutes. For example, the development of AGI poses an existential step risk; but, if there is a 'fast takeoff,' any additional state risk of developing AGI may be negligible.

Step risk versus state risk: How shall we determine when state risks are more important than step risks, or vice-versa? If a potential diminishment in a step risk depends on increasing a separate state risk (e.g. slowing down or stopping AGI research that, if successful, would decrease other risks to humanity), how do we decide what to do?

2.5 Acceptable-Risk Factors: Seriousness and Probability

We thereby come to the two most basic facets of risk assessment, seriousness and probability: how bad would the harm be, and how likely is it to happen?

Seriousness: A risk of death or serious physical (or psychological) harm is understandably seen differently than the risk of a scratch or a temporary power failure or slight monetary costs. But the attempt to make serious risks nonexistent may turn out to be prohibitively expensive or otherwise contraindicated. What magnitude of AI risk is acceptable—and to whom: users, nonusers, the environment, or the AI itself?

Probability: This is sometimes conflated with seriousness but is intellectually quite distinct. The seriousness of the risk of a 10-km asteroid hitting Earth is quite high (possible human extinction), but the probability is reassuringly low (though not zero, as perhaps the dinosaurs discovered). What is the probability of harm from AIs? How much certainty can we have in estimating this probability? How do we decide on the probability of serious harm that is acceptable, versus moderate harm or mild harm? If a function, is it linear, asymptotic, or other? Is it continuous or not?

2.6 Acceptable-Risk Factors: Who Determines Acceptable Risk?

In various other social contexts, all of the following have been defended as proper methods for determining that a risk is unacceptable [9]:

Good faith subjective standard: It is up to each individual as to whether an unacceptable risk exists. That would involve questions such as the following: Can the designers or users of AI be trusted to make wise choices about (un)acceptable risk? The idiosyncrasies of human risk aversion may make this standard impossible to defend, as well as the problem of involuntary/ non-voluntary risk borne by nonusers.

The reasonable-person standard: An unacceptable risk is simply what a fair, informed member of a relevant community believes to be an unacceptable risk. Can we substitute a professional code or some other basis for what a 'reasonable person' would think for the difficult-to-foresee vagaries of conditions in the rapidly emerging AI field, and the subjective judgment of its practitioners and users? Or what kind of judgment would we expect an autonomous AI to have—would we trust it to accurately determine and act upon the assessed risk? If not, then can AI never be deployed without teleoperators—like military robots, should we always demand a human in the loop? But even a 'kill switch' that enabled autonomous operation until a human doing remote surveillance determined something had gone wrong would still leave unsolved the first-generation problem.

Objective standard: An unacceptable risk requires evidence and/or expert testimony as to the reality of (and unacceptability of) the risk. But there remains the first-generation problem: how do we understand that something is an unacceptable risk unless some first generation has already endured and suffered from it? How else could we obtain convincing objective evidence?

2.7 Acceptable-Risk Factors: The Wild Card: Existential Risk?

Plausibly, a requirement for extensive, variegated, realistic, and exhaustive pre-deployment testing of AIs in virtual environments before they are used in actual human interactions could render many AI risks acceptable under the previous criteria. But one AI risk may remain unacceptable even with the most rigorous pre-deployment testing. An existential risk refers to a risk that, should it come to pass, would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential. Existential disasters would end human

civilization for all time to come. For utilitarians, existential risks are terribly important: doing what we can to mitigate even a small chance that humanity comes to an end may well be worth almost any cost. And for deontologists, the idea that 'one always has a moral obligation never to allow the extinction of all creatures capable of moral obligation' is at least a plausible *prima facie* (and perhaps absolute) duty; such a survival principle appears required for any viable ethics [10]. If there is even a tiny risk that developing AGI would pose an existential risk, this 'Extinction Principle' may well imply that we have a duty to stop it.

2.8 Conceptual Risk Analysis Matrix (CRAM^{SM})

Taking note of items 2.1-2.8 we have formed the following dialog matrix for focusing the dialog of AI Risks in a given use-case.

3. Specific Case Study, Possible Solution, and Risk Analysis

3.1 Specific Case Study

Our case study is applying CRAM^{SM} to the Path Planning of a Robot Arm, in which vials of severe biohazardous materials are to be moved from point A to point B in an optimum path. This path is constrained by parameters such as speed, power-usage, minimization of actuator acceleration and deceleration (that causes wear of the actuators) and collision avoidance. See Fig 1. Keep in mind that cost of production, as well as quality/safety, are value factors to be balanced in this manufacturing example. And the use of AI Robots in this case example is a very real-world example of potential benefit albeit with techno-ethical concerns. The engineering case study has been outlined in Refs [11], [12] and [13].

This path planning problem use case example envisaged here can be described in the following stages of "teaching" the system with a "show & tell" paradigm and AI-based optimization algorithm. It is a process of three stages teaching the system from a basic to advanced level novice to levels of expert proficiency.

Stage 1 proposes Mentor Training, whereby the motion capture of a human expert is captured into a simulation environment.

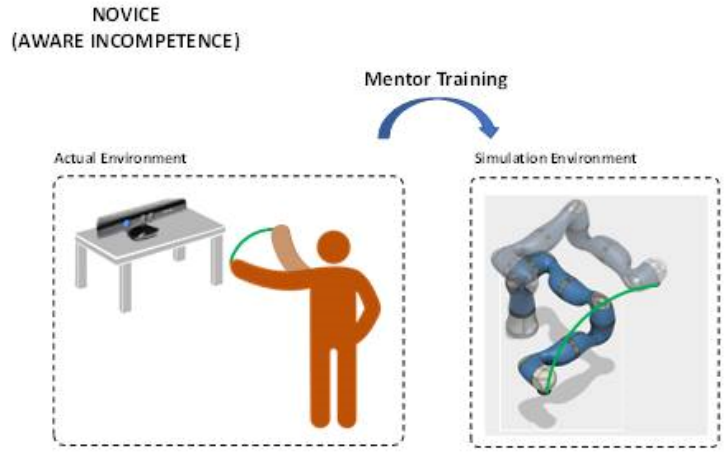
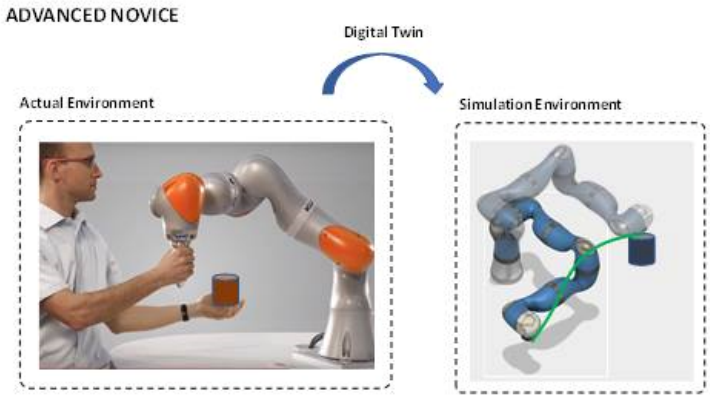


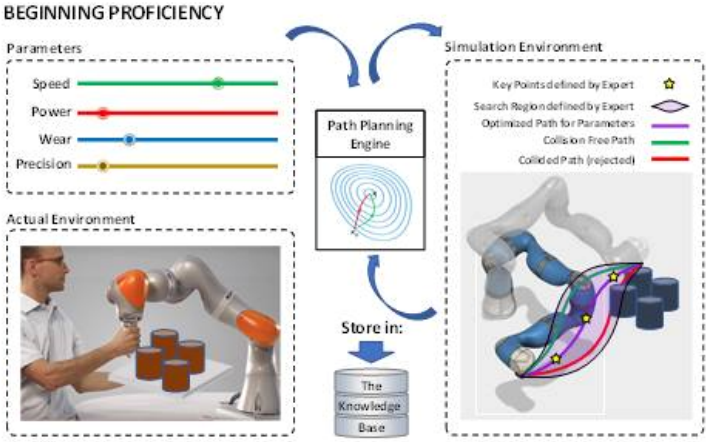
Table 1. Dialog of AI Risks in a given use-case

1/ Acceptable-Risk Factor: Consent
Assess the degree to which consent has been given of the use-case risk
2/ Acceptable-Risk Factor: Informed Consent
Assess the risk of the use-case that parties have potentially not been informed about (or do not understand) the risk potential
3/ Acceptable-Risk Factor: The Affected Population
Assess the risk to the potential Affected Population
4/ Acceptable-Risk Factor: Step risk versus State risk
Assess use case risk in terms of 1. *State Risk (time likely spent in a state that is a cause of risk)* 2. *Step Risk (chance of entering into a new risk, as a consequence of step transitions)*
5/ Acceptable-Risk Factors: Seriousness and Probability
Assess use case's risk in term of an analysis of potential seriousness and probability of occurrence 1. *Seriousness: What (if any) serious risks from AIs are acceptable—and to whom: users, nonusers, the environment, or the AI itself?* 2. *Probability: How much certainty can we have in estimating this probability? What probability of serious harm is acceptable? What probability of moderate harm is acceptable? What probability of mild harm is acceptable?*
6/ Acceptable-Risk Factors: Who Determines Acceptable Risk?
Assess use case's risk against standards applicable to it. 1. *Good faith subjective standard* 2. *The reasonable-person standard* 3. *Objective standard*
7/ Acceptable-Risk Factors: The Wild Card: Existential Risk?
Assess use case's existential risk that, should it come to pass, might 1. *annihilate Earth-originating intelligent life,* 2. *Or permanently or drastically curtail its potential.*

Stage 2 proposes Digital Twin capture of fine tuning of the Robot arm (a Co-roobotic arm is envisaged in this example that allows for the physical manipulation of the arm).



Stage 3 proposes an AI-Based fine tuning of the path planning by an optimization process based on value parameters (e.g., Speed of path cycle, Power utilization, Wear and Tear, and Precision of movement) an AI search algorithm might seek an optimum minimization of the path plan against a utility metric of these values.



3.2 CRAM[SM] applied to the Case Example

1/ Acceptable-Risk Factor: Consent
The use of a robot (in a protective clean room cell) reduces the need for human operator exposure to Biohazards. Any personnel entering the clean room cell should have safety training and contracted consent.
2/ Acceptable-Risk Factor: Informed Consent
However, if the AI directing the robot causes breaches of the clean/safe room (e.g., collisions with the cell walls), then what was thought safe might not be. In this respect personnel in the potential effective area may not be fully informed of the reliability/trust in the system. It is necessary to test any robot behavior in detailed simulation to ensure the path planning algorithms will not likely violate these rules. And personnel potential affected by failures be informed of the extent of the safety testing.
3/ Acceptable-Risk Factor: The Affected Population
The affected population might not be limited to the factory; conceivably, an extended exposure could cause health and safety threats to those outside, or violations of FDA regulations, etc. Again, the system must be validated against the regulations/laws applicable to the domain. The potential affected population should be briefed of the risks.
4/ Acceptable-Risk Factor: Step risk versus State risk
Both state and step risks need to be exposed through the AI testing process and the results passed to stage 5 below.
5/ Acceptable-Risk Factors: Seriousness and Probability
The seriousness of a biohazard breach can be evaluated in principle, but the probability may needs validating in test simulations. Thorough simulation is advocated (to avoid physical exposure) as well as an assessment of the system.
6/ Acceptable-Risk Factors: Who Determines Acceptable Risk?
There are industry bodies that set standards (e.g., GAMP5) as well as government entities that set regulations in this case example (e.g., US FDA).

7/ Acceptable-Risk Factors: The Wild Card: Existential Risk?
If the biohazard agent was severe enough, as might be possible with nuclear materials and/or live chemical/biological agents, then the impacts could be existential, if the AI goes "rogue." The severity relates properly to steps 3 and 5 above. The risk of 'going rogue' is conceivable, in a case of complete AI automation of the industrial facility, and absent proper safeguards against hacking. A solution may involve a software system over-riding ethical kernel that ensures "no harm" and sufficient cybersecurity measures. Ultimately a degree of human oversight may be warranted with the system requiring human authorization for critical procedures. Including the ability to switch off the system.

4. Conclusions

We reviewed the concept of AI Risk and picked a real world industrial problem. We proceeded to outline a means of structuring a dialog we refer to as CRAM[SM]

The AI/Robotics case example (AI-based Path Planning for a Pick and Place application for Biohazardous material) and applied the CRAM[SM] dialog to it; we think the result is an actual beneficial one for highlighting the AI risk concerns and start the process of handling them objectively.

As such, we believe the resulting techno-philosophy methodology to be a potentially useful early step in the building of tools for conceptualizing and assessing acceptable AI Risk. Further work is needed to develop these concepts, and trial them in real-world applications.

References

[1] UNIDIR: The weaponization of increasingly autonomous technologies: considering how Meaningful Human Control might move the discussion forward. UNIDIR Resources, no. 2, 2014. http://unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf. Last Referenced 22nd October 2017.

[2] Roff, H.: Monstermind or the doomsday machine? Autonomous cyberwarfare. *Duck of Minerva*, 13 August 2014. http://duckofminerva.com/2014/08/monstermind-or-the-doomsday-machine-autonomous-cyberwarfare.html. Last Referenced 22nd October 2017.

[3] Clevenger, A.: 'The Terminator conundrum': Pentagon weighs ethics of pairing deadly force, AI. *Army Times*, 23 January 2016. http://www.armytimes.com/story/defense/policy-budget/budget/2016/01/23/terminator-conundrum-pentagon-weighs-ethics-pairing-deadly-force-ai/79205722/. Last Referenced 22nd October 2017.

[4] Andrew Maynard, "Thinking innovatively about the risks of tech innovation". *The Conversation*, January 12, 2016. https://theconversation.com/thinking-innovatively-about-the-risks-of-tech-innovation-52934. Last Referenced 22nd October 2017.

[5] Lin, P., Mehlman, M., and Abney, K.: Enhanced warfighters: risk, ethics, and policy. Report funded by the Greenwall Foundation. California Polytechnic State University, San Luis Obispo. 1 January 2013. http://ethics.calpoly.edu/Greenwall_report.pdf. Last Referenced 22nd October 2017.

[6] Abney, K., Lin, P., and Mehlman, M. "Military Neuroenhancement and Risk Assessment" in James Giordano (ed.), *Neuroscience and Neurotechnology in National Security and Defense: Practical Considerations, Ethical Concerns* (Taylor & Francis Group, 2014)

[7] Lin, P., Mehlman, M., and Abney, K.: Enhanced warfighters: risk, ethics, and policy. Report funded by the Greenwall Foundation. California Polytechnic State University, San Luis Obispo. 1 January 2013. http://ethics.calpoly.edu/Greenwall_report.pdf . Last Referenced 22nd October 2017.

[8] Nick Bostrom, *Superintelligence.* (Oxford University Press, 2014)

[9] Abney, K., Lin, P., and Mehlman, M. "Military Neuroenhancement and Risk Assessment" in James Giordano (ed.), *Neuroscience and Neurotechnology in National Security and Defense: Practical Considerations, Ethical Concerns* (Taylor & Francis Group, 2014)

[10] Keith Abney, "Robots and Space Ethics," ch 23 in *Robot Ethics 2.0*, eds. Lin, P., Jenkins, R., and Abney, K. (Oxford University Press, 2017)

[11] M. Ciupa, "Is AI in Jeopardy? The Need to Under Promise and Over Deliver - The Case for Really Useful Machine Learning," in *Computer Science & Information Technology (CS & IT)*, 2017.

[12] M Ciupa, N Tedesco, and M Ghobadi, “Automating Automation: Master Mentoring Process” 5th International Conference on Artificial Intelligence and Applications (AIAP-2018), Jan 2018, Zurich, Switzerland
[13] M Ciupa and K Abney, “Conceptualizing AI risk” (AIFU-2018), Melbourne, Australia February, 2018.

ICRES 2018: International Conference on Robot Ethics
and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.024

JANUS-HEADED ROBOTICS: DILEMMAS AND PARADOXES IN ROBOT ETHICS

ENDRE E. KADAR
Department of Psychology, University of Portsmouth, Portsmouth, PO1 2DY, UK

In designing robots with high level of autonomy for safe and ethically acceptable interaction with humans, engineers need to have a profound understanding of human behaviour control including social interaction. The present paper was inspired by Janus, a god of transition in Roman mythology whose double face could be used as a metaphor to provide insights into current difficulties and ethical concerns of designing artificial autonomous agents. A critical review of biology and the theoretical background of ethical concerns in biology are used in order to gain insights in recent developments in robotic research and ethical concerns in the context of the history of Western Science. Aristotelian organic science was contrasted with the Newtonian-Cartesian science to discuss the dilemmas and paradoxes of robot ethics arising from insufficient understanding of the autonomy of an agency. These are presented from the two perspectives Janus head is facing: looking backwards to see the dangers and benefits of past and current limitations of our knowledge in robot designs; and looking forward benefitting from or abusing of our future deeper understanding of the principles of normal functioning of autonomous agents.

1. Introduction

In designing robots whose interaction with humans is safe and ethically acceptable, engineers need to understand human perception and action in behaviour control including those skills that are needed in successful interaction processes [1]. In other words, fluent human-robot interaction requires profound understanding of human social and non-social interaction skills. Biology provides examples of systems with complex action repertoire and high level of autonomy. In contrast, designing artificial systems with high level of autonomy is still a challenging task even though the field of robotics is developing fast. To enhance safety, one of the common strategies is to develop mixed control system that allows human interference if the robotic control system is malfunctioning. An additional advantage of the mixed control design is that human performance could also be monitored and human errors can be corrected. But this mixed control systems have their challenges too. They also imply having finely calibrated basic interaction skills because the performance of an automatic controller should be similar to human-like behaviour otherwise the human controller would have difficulties in detecting the need to take over the control when a possible error/malfunctioning of the automatic control mechanism occurs. Similarly, the

automatic controller (artificial agent) should be able to monitor human performance in order to warn the human agent and/or take over the control when obvious human errors are detected.

The present paper was inspired by Janus, a god from ancient Roman mythology, whose metaphoric role could provide insights into current difficulties and ethical concerns of designing artificial autonomous agents. Janus was the god of transition, beginnings, doorways, passages and endings. He is depicted with two faces, but the original meaning of these two faces is different than contemporary interpretation of the term "Janus-headed". The two faces are traditionally representing the two directions of looking, that is, looking into the past and future rather than having a double identity, the common negative connotation of the contemporary interpretation. But Janus was also associated with a different negative meaning in ancient Rome. He was the god who presided over conflicts from the beginning to ending. These traditional meanings are useful to illustrate the state of the art in contemporary robotics. On the one hand, looking back at the problems and limitations of past robotics to learn valuable lessons from the past. On the other hand, we can look forward and try to anticipate future challenges to be prepared to overcome foreseeable difficulties. To put differently, we live in a transitional period of the history of robotics. The past few decades were dominated by intellectual efforts to overcome various limitations of our knowledge and technology. These limitations allowed the use of limited autonomy with limited danger only but the new era of robotics will be dominated by increasing level of autonomy together with increasing levels of both benefits and dangers.

The discipline of robot ethics has emerged as a new area of research to mark this transitional period with lots of foreseeable and unknown dangers, sources of fears and anxieties in contemporary societies. The research of robot ethics and safety can be regarded as a battleground for opposing forces of development such as the desire of using autonomous designs and the fear of losing the ultimate human control over all robot systems. This situation, however, is not a new phenomenon in the history of modern sciences. For instance, a century ago, Physics had a similar period at the birth of quantum physics when the new insights into nuclear forces created the fear of the danger of abusing this knowledge that could lead to a nuclear catastrophe. Similarly, a few decades ago, Biology entered a similar transitional period when genetics research opened up new possibilities and at the same time became a source of various concerns. Fabricating new life forms with genetic engineering such as genetic manipulation can be used to treat genetic disorders but can also be abused in many different ways. New life forms are new agents with not known or not fully understood behaviour patterns. Even a simple system such as a bacterium or a virus could be modified in such a way that they could be used as biological weapons to kill huge number of humans. These biological concerns are very similar to the problems of new robotics, because contemporary researchers are able to design robots with much higher-level autonomy than in the past. The obvious similarity of fabricating new life forms and fabricating autonomous robots has inspired the present paper to learn valuable lessons. Thus, we first look at how biology was influenced by the tradition of Western Science, and how

the machine metaphor constrained its developments while a new biology (including bionics, genomics, etc.) has emerged. Then, our paper highlights that the prospect of a radically new approach to biology outlined by Rosen [2] can be considered as an analogue to the age of new robotics dominated by principles of autonomous agency building on Rosen's [2,3,4,5,6] insights. Janus is called again to preside over "another war", an "intellectual war" between the old tradition and a new emerging field of research, robotics of autonomous agents. The paper concludes with the presentation of a few dilemmas and paradoxes to consider that help overcome the limitations of old robotics and facilitate the birth of new robotics.

2. The Problem of Agency in Science

The birth of Western Science is usually associated with Aristotle whose physics (Natural Philosophy) and metaphysics laid its foundation. His view on science, however, was profoundly different than modern science that is based on the works of Newton, Galileo, Bacon and various other prominent thinkers of modern era [7,8,9]. Artistotle's Science (Natural Philosophy) is often labeled as organic based on the fact that motion was its central concept and each motion is associated with a mover (Bk. 8 of the *Physics* argues for the additional thesis that for each motion, whether natural or contrary to nature, there needs to exist a mover.) To put differently, within this system agency is central because the notion of mover could be used to define agency, which includes the first mover (God), living creatures and could also be artificial mechanisms (e.g., tools, machines, etc.). Another important part of Aristotle's science is the way things, changes, non-changes in Nature are described to provide explanations of changes and non-changes. To be more exact, the famous doctrine of four causes is meant to provide explanation of why things are as they are or why changes happen the way they do. These four causes can be grouped into two pairs of causes. The more fundamental pair of causes is the material and formal causes. Material causes provide potentials, which are or can be actualized by the formal cause. Thus, these two are related rather than independent aspects of explanations. The other two causes are more closely related to changes in Nature. The efficient and final causes are beyond but not necessarily independent of the material and formal causes. The efficient causes initiate processes that lead to changes, while the final cause is the end of the process of changes, what efficient causes intended to achieve. Aristotelian approach could provide a relatively simple framework for the theoretical issues surrounding the problem of agency in robotics, in particular autonomous robot designs. Specifically, the chain of causes can be infinite but Aristotle emphasized that often this chain has a beginning, and that is associated with an agent (god, animal, human, etc.), which is a primary cause of the unfolding changes over a period of time [7].

Some of Aristotle's ideas are still appreciated but modern science developed on a radically different basis. Philosophers (Spinoza, Descartes, Bacon) and scientists

(Newton, Galileo) during the Renaissance and ensuing centuries challenged some of the assumptions of Aristotelian science [8,9]. Specifically, Newton was keen on eliminating the agency from Nature by elaborating the foundation of Natural Philosophy on the basis of mathematics and laws that do not require or imply mover (agency) in explaining how things are and why changes occur. Newton developed classical mechanics, which does no have an agency/mover but Newton's ambition of creating a systematic approach that could explain everything in Nature failed (e.g., gravitational attraction remained unexplained). Despite its limitation, the Newtonian classical mechanics became so popular that the world-view of the universe as a huge mechanism became widely accepted. Accordingly, quite often the machine metaphor was used more specifically and the universe is likened to gigantic clockwork mechanism. This view is deterministic in a limited sense with regard to Aristotle's four causes. In the Newtonian universe, the efficient cause became dominant and the material and formal causes of mechanisms remained still useful but played far more limited role than in the Aristotelian system. The "spooky" final cause that is mostly associated with an agency became meaningless and unnecessary because it was regarded as either an illusion or something that can be reduced to other causes or something that can be eliminated altogether. The role of agency as a primary cause was eliminated because it was mostly reduced to a specific version of efficient cause.

Descartes [9] was a strong supporter of this mechanistic view, but interestingly, he realized that humans do not fit into this universal clock-work mechanism. He postulated that humans have free will that is essential ingredient of his dualist view of humans who consist of matter and another substance that constitutes the mind. To put differently, human mind is different to matter, because the bodily machinery is void of agency but the mind can play the role of an agency. Thus, Descartes, perhaps not intentionally, brought back Aristotle's agency into modern science. This mind-body dualism survived centuries and it is still popular within Psychology (in particular, cognitive psychology promotes, for instance, the theory of mind, etc.). Two important aspects of the Cartesian view, however, have changed dramatically. One of these was the completion of the mechanistic view of the world that includes mental aspects. This was due to development in language and logic about a century ago, which led to the computational theory of mind that brought human (animal) mind back into the Newtonian universe by making the mind computational (a specific kind of mechanism) [10]. The other important trend is that scientists no longer believe that mental life is unique to humans. Most scientists accept the view that has already been anticipated by Darwin, who claimed that mental ability is not unique to humans and argued for the continuity of mental life in evolution. Darwin [11] published a famous research on emotion that clearly challenged the Cartesian view, which postulated that animals are mindless and non-sentient beings. He suggested that many other non-human species have mind, which could be different for different species but in many ways similar to human mind.

These two changes have important impact on the notion of agency in biology, psychology and robotics. Across these disciplines, two opposing and competing trends

emerged during the past 150 years. The dominant trend (mainstream view) is to accept the scientific stance that living creatures and their mental life can be viewed as part of the mechanist universe. In other words, agencies can be described and their functioning can be explained as fundamentally mechanistic even though in many ways they are simply a specific kind of mechanistic systems. The other, opposing view of agency in these three disciplines is to separate living creatures from the reductionist mechanistic view of the universe. The separation can be made on an ontological basis similar to the Cartesian dualism, except the dualism would be between non-living part of the universe and living creatures. But there are other possible separations such as differentiating organisms with or without nervous systems, organisms with or without consciousness, etc. It will be argued in this paper that there is a fundamental similarity the way biology was developed and raised ethical concerns several decades earlier before similar developments raised similar concerns in robotics. This is the motivation for reviewing the notion of agency and associated ethical concerns in biology.

3. The Problem of Function and Agency in Biology

3.1. *Vitalism, Organismic Biology and Relational Biology*

The mystery of life, what makes living creatures different to non-living things is a very old problem. It was already discussed in ancient texts, including the works of Aristotle. This problem was framed in different ways as human knowledge developed through the ages. During the 18th and 19th centuries there were two alternative approaches: 1) the Cartesian mechanistic reductionist explanations of living organisms as an agency based on non-living matter; and 2) the vitalist view that postulated additional extra material/substance or principle required for life. Bechter and Richardson [12] summarizes vitalism:

> According to vitalists, living organisms are fundamentally different from non-living entities because they contain some non-physical element or are governed by different principles than are inanimate things. Various forms of vitalism have been developed. Some argued that living entities contain some fluid, or a distinctive 'spirit'. Other, more sophisticated versions promoted the idea of the vital spirit becomes a substance infusing bodies and giving life to them, etc. Modern vitalist views were developed to oppose Cartesian mechanistic view of organisms and there were prominent advocates of this view early as recently as in the twentieth century (e.g., Hans Driesch (1867–1941), an eminent embryologist, argued for the presence of an *entelechy*, a substance that controls organic processes, Henri Bergson (1874–1948) argued for *élan vital* to overcome the resistance of inert matter).

Organismic biology was another approach that found both the Cartesian mechanical explanation and vitalism unsatisfactory in searching for an alternative explanation of the unity of living organisms. Ritter [13] is often considered as the father of this view but some would argue that Aristotle was not a vitalist and his approach was the first

version of organismic biology.

Ritter [13] argued that the totality of the organism is as essential to an explanation of the behavior of its elements as elements are to an explanation of the behavior of the organism. Woodger's [14] *Biological Principles* attempted to explain organismic unity via principles of hierarchical organization. At the lowest level, elementary physical particles can be described by physical laws. According to Woodger, a system is perfectly organized hierarchically if parts of one level are the sole constituents of the next higher level. Woodger [14] considered organisms to be fundamentally hierarchical, even though there are exceptions such as the cardiovascular system, which is a subsystem that is not hierarchically organized. Although its tributaries are hierarchically organized over the size dimension, ranging from major arterial and venal pathways to tiny capillaries but is a more-or-less closed loop system. Whilst it is true that not all organisms possess cardiovascular subsystem (e.g., obvious exceptions are protozoa, bacteria, etc.), this subsystem is closely linked to the fundamental metabolic processes that are characteristic of all organisms.

Rashevsky [15] has taken a different approach. He checked various physical principles used in the literature and reviewed all known attempts to capture the "essence of life" and he realized that all of these efforts were futile. He found the cardiovascular subsystem useful to gain insights into strange circular processes in living system. Rashevsky [15] developed an argument to capture the essential features of living organisms based on the fundamental processes. Mapping the interrelationships of all basic processes in graph theoretic terms he no longer had to rely on cardiovascular system and he could propose a radically new approach that could accommodate plasticity (i.e., differences in species as well as individual differences in organisms) as well as circularity in fundamental biological processes.

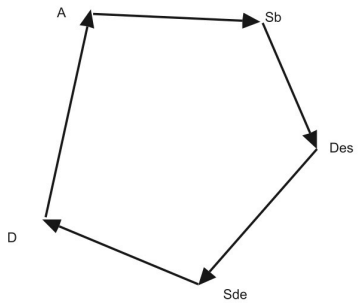


Figure 1. The common cyclic process of the relational structure of living organisms in Rashevsky's [15] study.

Rashevsky [15] decided to look at functional processes in a variety of organisms. Representing these basic processes by transformations using directed graph (digraph) techniques, he was able to develop a topological representation of

the organism as a functioning whole for a specific species. Figure 1 shows how the feeding process can be represented by a directed graph. Feeding begins with ingestion followed by digestion (D) and, in Rashevsky's [15] words:

> This is followed by the absorption *A* of the digested food into the protoplasm of the cell, as well as by the rejection of indigestible waste, or defecation *Df.* The absorbed food is transported to various parts of the cell where a synthesis S_b of the body of cell follows. These general synthetic processes result, among other things, in the synthesis D_{es} of digestive enzymes. This is followed by the secretion S_{de} of the digestive enzymes into the digestive vacuole, where they come into contact with the ingested food, and digestion *D* results. The directed path $IDAS_bD_{es}S_{de}D$, which contains the cycle $DAS_bD_{es}S_{de}D$, expresses a very fundamental property of every organism: in order that food may be assimilated and the body of the organism synthesized, the organism must be already present (pp. 326–327).

In sum, living organisms are shown to be different to non-living entities because they include a fundamental metabolic processing system that is common across all species. This fact could be framed by using Aristotle's notion of causality. What makes biological agents different? How biological agents maintain their lives? Why do living creatures differ to non-living entities? These and many other similar questions could be raised and the answers would include causations of various kinds that are linked to or based on the fundamental cycle revealed by Rashevsky [15]. This insight could provide the foundation of looking at the problem of autonomous agency in a new way.

3.2. *Rosen's Complex System Approach*

Rashevsky's student, Robert Rosen realized that the relational structure is actually a simplified form of more abstract mathematical structures, complex chain of mappings that are derived in category theory. This conclusion was based on a number of important steps in modeling and understanding living organisms, but the gist of his approach can be explained in simple terms. Rosen [2] noticed that biology is dominated by the mechanistic (Newtonian-Cartesian) view of the world that could only provide simple reductionist models that are inadequate to capture fundamental aspects of living organisms. He phrased the problem simply: What makes an organism different than a machine? Rosen's answer is seemingly simple: Machines are simple and organisms are complex. Machines are simple because they can be taken apart and they could also be put together from its pieces. Complex systems can be defined in contrast with simple ones. All systems are complex if they are not simple (i.e., the system is more than simply a some of its parts). Many of the physical systems that are known as simple mechanisms, in

fact, are complex. Organisms are all complex systems. This intuitive definition of complexity can be controversial because there are many other definitions of complexity, especially if the definition is based on the behavior pattern of the system. For sake of simplicity, the present paper uses Rosen's intuitive definition by contrasting complexity with simplicity and equating simplicity with mechanisms (machine metaphor).

Rosen has arrived at this position based on a critical review of modern science. He realized that Newtonian – Cartesian science lost the appreciation of Aristotle's four causes partly because the focus of modern science is on answering "how" rather than "why" questions. Even if the scientific analysis involves causes, the final cause is mostly missing or reduced to the other three. But final cause is clearly important in organisms, which are goal directed anticipatory systems [5]. Rosen argued that complex systems always involve all four causes that are mixed (interlinked) and cannot be separated as they can in machines.

The second important aspect of Rosen's approach is to focus on the functional process rather than structural components of the system, which is typical of machine-based views of the world. He realized that proper handling of time is a problem in dynamics, but he side-stepped this issue by using mappings. Mapping was also instrumental in revealing complex relationships based on the loops generated by mappings. With mappings, closed loops can emerge that would be similar to Rashevsky's closed loop in Figure 1.

Thirdly, following Rashevsky [15], Rosen shifted the focus from structural components typical of machines to functional components of complex systems such as organisms. He specifically investigated the problem of how to repair malfunctioning and noted that functional components could and should be replicated and this replication should be done by itself [3]. In other words, the replication will be self-replication. This issue of self-referential element in the system is an enigma for standard science (please note that Russell's effort to eliminate paradoxes from science resulted in elimination of self-referential statements from mathematical/formal logic of science). Rosen has developed a technique that addresses the problem at the proper level (high level of abstractness).

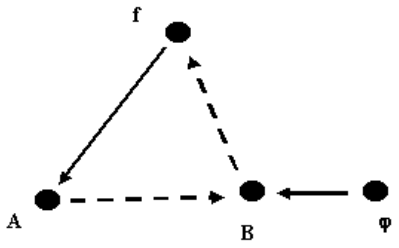
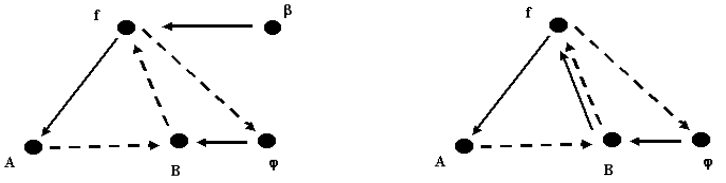


Figure 2. The broken lined arrows represent material causation and the solid arrows represent efficient cause. If we note that f has the meaning of some *relation* on the Cartesian product of sets A and B [H (A,B)]. This can also be written as a mapping: f φ A → B → H(A,B).

In Rosen's approach, basic functional issues are modeled as relational processes that can be represented as a mapping diagram. For instance, f, represented as f: A → B. A can be considered a set of initial conditions, B a set of final conditions, and f denotes the particular ways that some change was brought about to transform A to B. Rosen then assigns to this symbolism a set of entailments using the Aristotelian "why?" question. For example, to answer the question "why B?" there are at least two answers. A is the material cause of B. The mapping, f, is the efficient cause. We might also ask "why f?" and the answer would be "to bring A into B" assigning to f a final cause role.

Rosen tried and add repair function to the system represented in the mapping structure. This mapping, however, is just a simple one by adding another φ mapping component (see Figure 2) that could provide the cause of f, but this mapping could represent a simple machine that requires external interference (action). This is because the cause of φ is not going to be part of the system itself. By introducing another mapping to take care of the cause of φ, one could add another cause ß and one could continue adding another cause to end up in an infinite regress [3]. Figure 3 demonstrates how this infinite regress could be avoided by creating a complex system that does not include external causes. To put differently, to eliminate infinite regress of causes, all causes should be made intrinsic rather than external. This would mean that the diagram would no longer have arrows that imply causes that are extrinsic (in other words, they are not entailed within the system if we use Rosen's terminology). This could also be expressed as a way to show how autonomy could be represented/defined with regard to a specific function. All functions could be used to achieve enhanced autonomy, which could be modeled by Rosen's mapping technique.



(a) (b)

Figure 3. In a machine-based approach, diagram (a) shows that we can added the replication step ß to entail (add a cause for) φ but it would have left ß "unentailed" (not caused internally), for given that ß is really ß: $f \rightarrow \varphi$, that is that it is the efficient cause of φ and that f is the material cause of φ, ß is left "unentailed". If it is established that it is possible that $ß = b^{*-1}$, the inverse of $b^* : f \rightarrow B$, so that the diagram can be redrawn (b) to ensure that no external cause is necessary for repair function (i.e., repair became self-repair).

3.3. *From Biology to Bionics and Ethical Concerns*

Rosen was worried about the problems, including the danger and ethical concerns biological researchers may cause. Those concerns included bionics research (the study of *mechanical systems* that function like living organisms or parts of living organisms) raised contemporary societies were mostly worried. To put it differently, he was aware of the danger of fabrication and manipulation of life, but he was not deeply concerned because he argued that as long as machine-based thinking dominates bionics (*nomen est omen*), there is no real danger of causing major problems. Bionics is also defined as biologically inspired engineering, that is to say, this discipline is also associated with applying biological methods and systems in engineering. If bionics becomes a truly biologically inspired engineering, that is using insights derived by methods Rosen was working with, that will be alarming.

In our discussion of Rosen's insights and theoretical works on complex systems, including organisms, the word autonomy was not central notion because that was not part of Rosen's jargon in discussing his methodological and theoretical considerations. But it is obvious, that his notion of complexity implies that autonomous systems are complex. To put differently, without having a complex system embodied in a robot, the robot remains a simple machine only. Thus, recent concerns regarding the dramatic increase in the level of autonomy of contemporary robot systems can be directly linked to Rosen's theory and can be discussed rigorously within his framework.

4. Lessons from Biology for Robotics and Robot Ethics

Clearly, robotics was and still is a discipline that is part of the mainstream of modern science. It is well-known that robotics research increasingly relied on biologically inspired performance during the past few decades. So what this fuss is about? Robotics researchers could be and are indeed often misguided by the fact that biology and psychology are also victims of Newtonian-Cartesian scientific tradition. When researchers are looking for insights in biology or psychology, they may find that mainstream views can be accommodated in robotics. But they do not realize that those methods and insights are not truly biological and natural, rather they are simplistic due to the limitations of mainstream science (mechanistic models). Specifically, when researchers are talking about autonomy in designing robotic system, they mean a fairly limited autonomy that would still leave the system a machine, that is a simple system rather than a complex one using Rosen's definition. This is due to the fact that robots are mostly dependent on external control. Also, the "internal" control mechanisms are not truly internal because they are part of the software that is either rigidly hard-wired (burnt in) or external to the hardware. Minimally, designing autonomous robots based on Rosen's conceptualization would require analogue systems, rather than the commonly used hardware-software digital systems that are based on the Cartesian dualist mechanistic tradition. To put it differently, most contemporary robot systems are the results of a historic continuity in robotics that can be summarize in a brief overview.

Early robots were designed based on clockwork-driven smart mechanisms but remained "mindless" systems until language processing became available with the help of mathematics and computers. In other words, with the advent of computers both mind and body became part of a mechanism and both are physical and computational. In the 1960s and 70s, the cognitive revolution in psychology was the byproduct of this development. Accordingly, the body of the agent and its surroundings are represented symbolically in the modular structure of the agent's Cartesian "mind". Information about the body of the agent and its environment is processed based on sensory data, movement plans are designed and executed based on the output of motor control (executive) modules. The idea of modular structure of the mind is also derived from the machine metaphor of mind, which is integral part of the Cartesian-Newtonian tradition of science.

During the past 50 years, artificial intelligence and robotics research adopted this cognitive architecture as evidenced by basic textbooks [16]. Based on our critical review of biology and Rosen's work, it should be clear that this cognitive architecture-based model of robotics still has major limitations. It is based on mainstream psychology that remained the prisoner of the Newtonian-Cartesian view of science, the science of mechanisms and machines but this approach has received several critical comments by prominent researchers. Brooks [17], for instance, noted that there is no need to create representations because the environment can itself be used for computations. Representational approaches are operating within the Cartesian tradition whereby the physical body (hardware) and

environment have a dual presence in the mind-software, which is extrinsic to the hardware and the physical environment. Others argued that Gibson's [18, 19] radical theory of perception should be used to eliminate representations. This approach had a few additional advantages including its intrinsic dynamic nature having the close link between perception and action by breaking the rigid separation of sensory and action control systems. These ideas could be adopted in robotics research [20]. Gibson's [19] theory could be helpful in paving the way towards a new dynamic and functional robotics that is fundamental in developing autonomous robots.

5. Conclusions

The present study is based on a critical review of biology and the theoretical background of ethical concerns in biology in order to gain insights in recent developments in robotic research and ethical concerns. Our review of recent developments in biology and robotics was situated in the context of the history of western science. Aristotelian organic science was contrasted with the Newtonian-Cartesian science that dominated the past few centuries. This comparison was based on two important aspects: The Aristotelian four causes and the problem of agency. Using Rosen's [2,3] theoretical work in biology, the importance of Aristotelian four causes were highlighted and category theory was introduced to demonstrate how abstract problems can be discussed in a formally rigorous way.

There are many studies on the problem of comparing natural and artificial agents with their intelligence. For instance, Dreyfus [10.21] published his philosophical studies in support of his arguments on the limitations of artificial intelligence. The author of the present paper has also argued in support of fundamental limitations of robotics based on the fact that engineers are not taking into consideration the difference between the life world and the abstract world of science [1,22]. However, these studies and their arguments are not sufficiently effective because of the lack of proper formalism that is suitable to discuss these abstract problems. The present paper tried to remedy this shortcoming and argued that Rosen's theoretical framework with the formalism of category theory is suitable to discuss agency at an abstract level. Mappings are representing various types of Aristotelian causes in Rosen's discussion of complex systems of biological organisms. The diagrams represent agencies that absorb external causes and driven by internalised causes with minimal external forcing (high degree of freedom, not dependence on external effect – high level of autonomy by definition).

The sources of fundamental ethical problem of bioethics/bionics and robot ethics/robotics are very similar (i.e., avoiding the creation/fabrication of agencies that could pose significant threat for humans). To protect humans from being victimized by the adverse affects of recent developments of biology or robotics similar ethical concerns can be associated with artificially created agents. It is

argued that the complexity of autonomy and associated dangers can be discussed with the help of Rosen's mathematical modeling strategy.

Overall, the state of the art in robotics can be described by a Janus head whose faces are directed towards the past and future. Looking backward we can still feel the presence of the burden of the inherited limitations of the machine universe of the tradition of modern Western Science. But if we look into the future, we can see the lurking dangers associated with creating robots with increasing level of autonomy. The apparent paradox of robot ethics is a new version of the paradox of knowledge, which is one of the well-known old paradoxes of human condition. Ignorance is bliss and knowledge is power, but humans can suffer from the consequences of both ignorance and knowledge. Our goal is not to benefit from ignorance like swindlers tend to do by relying false promises rather we should learn the lessons from the failings of the past, so we can be looking forward to benefitting from robot autonomy while facing an coping with the challenges they bring with them.

References

1. 1.E. E. Kadar, A. Koszeghy, and G. S. Virk, Safety and ethical concerns in Mixed Human-Robot Control of Vehicles. In M. I. S. Ferreira, J. S. Sequeira, M. O. Tokhi, E. E. Kadar, G.S. Virk, *A World with Robots: International Conference on Robot Ethics: ICRE 2015*. (pp. 135-144). Springer. (2017).
2. 2. R. Rosen, *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*. New York: Columbia University Press. (1991).
3. 3. R. Rosen, Some relational cell models: The metabolism-repair systems, in *Foundations of Mathematical Biology, Volume II,* R. Rosen, ed. Academic Press, NY. (1972).
4. 4. R. Rosen, *Fundamentals of Measurement and representation of Natural Systems*. New York: North-Holland. (1978).
5. 5. R. Rosen, *Anticipatory Systems: Philosophical, Mathematical & Methodological Foundations*, New York: Pergamon Press. (1985).
6. 6. R. Rosen, On the limits of Scientific knowledge in *Boundaries and barriers:on the limits to scientific knowledge.* (J. L. Casti and A. Karlqvist, eds.). pp. 199-214. Reading: Addison-Wesley. (1996).
7. 7. Aristotle, *The Complete Works of Aristotle.* J, Barnes (Ed.). Blackwell. (1984).
8. 8. A. Stinner, The Story of force: from Aristotle to Einstein. *Phys. Educ.,* 29 77, (1994).
9. 9. D. Garber, Descartes, Mechanics, and the Mechanical Philosophy. *Midwest Studies in Philosophy*, 25, (2002).
10. 10. H. L. Dreyfus, What computers can't do: The limits of artificial intelligence. (1972).
11. 11. C. Darwin, The expression of the emotions in man and animals. (1872).
12. 12. W. Bechtel, and R. C. Richardson, Vitalism. In E. Craig (Ed.), *Routledge Encyclopedia of Philosophy*. London: Routledge. (1998).

13. 13. W. E. Ritter, *The unity of organism.* Boston: R.G. Badger. (1919).
14. 14. J. H. Woodger, *Biological principles, a critical study.* London: Kegan Paul. (1929).
15. 15. N. Rashevsky, "Topology and life: In search of general mathematical principles in biology and sociology". *Bull.Math. Biophys.* Vol.16: 317-348. (1954).
16. 16. B. Siciliano and O. Khatib, *Springer handbook of robotics.* Springer. (2016).
17. 17. R. A. Brooks, Intelligence without representation. *Artificial Intelligence*, *47*(1), 139-159. (1991).
18. 18. J. J. Gibson, *The senses considered as perceptual systems*. Boston: Houghton Mifflin. (1966).
19. 19. J. J. Gibson, *The ecological approach to visual perception.* Original work published 1979. New Jersey: Lawrence Erlbaum Associates. (1979/1986).
20. 20. A. P. Duchon, W.H. Warren and L.P. Kaelbling, Ecological robotics. *Adaptive Behavior, 6*, 473-507. (1998).
21. 21. H. L. Dreyfus, What computers still can't do: A critique of artificial reason. MIT Press. (1992).
22. 22. E.E. Kadar. Mind the gap: A theory is needed to bridge the gap between the human skills and self-driving cars. In press.

ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.003

WHY DO WE NEED ROBOTIC & AI GOVERNANCE? AN ANALYSIS OF THE (SOCIO-) ECONOMIC IMPLICATIONS OF ROBOTICS AND ARTIFICIAL INTELLIGENCE

DOMINIK B. O. BOESL

Technical University Munich (TUM), Munich, Germany; Vice President IEEE RAS Industrial Activities; Chair IEEE TechEthics ad-hoc Committee; Senior Innovation Manager & Vice President Consumer Driven Robotics with KUKA AG, Augsburg, Germany

MARTINA BODE

Junior Research Manager with KUKA AG, Augsburg, Germany

This paper illustrates the necessity of a Code of Conduct for Robotics and Artificial Intelligence – ethical and moral guidelines for the development and use of these technologies – by analyzing some of the potential (socio-) economic effects that might arise from technological advances in these fields. Contrary to other works in the field of roboethics, the authors did not strive to analyze individual cases that are of interests for ethicists and moral philosophers, but want to offer a broad, holistic and economic view on the effects of robotics and artificial intelligence on our future economy and, in the second instance therefore also society, as well as on the concept of Robotic & AI Governance for self-regulation of ethical, moral, sociocultural, socio-political and socio-economic questions, using a market mechanism.

1. Introduction

Robots are currently changing not only our manufacturing processes, but are gaining importance in other fields like service or consumer/household robotics. Europe's public view on these developments is rather critical and the fear that robots could replace jobs in the future is still present, even though studies suggest, that this is not the case (e.g. [1]) and that robots are currently creating more jobs than they are replacing. The European governments on the other hand are investing heavily in new technologies like service robotics, to create a competitive advantage for the region in the future. In countries like Japan, for example, robots are more accepted by the public, in general, due to cultural reasons.

Due to these changes, ethical and moral standards have to adapt. Researchers and scientists should critically ask themselves how their inventions will be used, who will profit from them and who will not as well as if they are sustainable [2] or if they can be misused, e.g. to commit crimes. In general, all these questions can be broken down to some general statements regarding for example the freedom of science and if it justifies the uncontrolled development of new and potentially dangerous technologies. To restrict research and development, on the other hand, means to hinder innovation.

Another one of these points is the missing "humanness" of machines. They cannot feel anything like compassion or sympathy and therefore it could be necessary to restrict the use of robots to fields where that lack of these qualities cannot lead to ethical conflicts. Asimov's laws [3] of robotic for example do not allow the robot to harm humans, but is not considering this specific question regarding the lack of empathy. This leads to the idea if it could be possible to implant such a system of values into a robot. And could these values be compatible with Asimov's laws? Many use cases, e.g. in service robotics would potentially imply a conflict of interest between the human's freedom of choice and Asimov's laws, for example if a robot is asked to serve alcoholic beverages to a customer, which are generally known to be harmful to health. How can a robot be enabled to make that decision? Can it distinguish between customers that are capable of taking that decision on their own (adults) and customers that are not (e.g. minors or mentally handicapped people)?

All these questions illustrate how complex the discussion regarding roboethics is. The goal has to be to prepare the generation of 'Robotic Natives' [4] to handle those autonomous robotic agents, to benefit from them and to co-exist with them.

2. Effects on the Labor Market

Thinking about the future of human labor, it has to be examined in which cases humans can, might or even should be replaced by machines and in which this seems impossible, improbable and undesirable. Many researchers think that the idea of an independent robot is science fiction. At the moment, automation of work is mostly restricted to industrial robotics, housed in factories and heavily constrained environments. The outside world is unpredictable and robots are in general better at pre-defined and narrow tasks and contexts. In addition, there are many legal questions yet unsolved [6].

Therefore, complex cognitive tasks, creative work as well as social-emotional intelligence can and will not be replaced by robots in the near future. Human labor will most likely experience a shift from physical work to more creative and complex jobs like artists, designers or public relation specialists. In addition, new, creative jobs might evolve in this process. The job market of the future will also see more part-time jobs as well as jobs where robots and humans work together. In addition, it is reasonable to expect that more people will work in jobs where social skills are important, such as e.g. elderly care, which could in addition help to solve the problem of labor shortage in these fields [7].

There are many studies regarding the amount of jobs that can be automated in general. Nevertheless, the utilization of new technologies is slowed down by economic, legal and social hurdles and the robotics industry is also creating new jobs. Taking heterogeneity of workers' tasks within occupations into account, across the 21 OECD countries, on average 9% of jobs are automatable, according to a report by Arntz et al. (2016) [8]. On the other hand, it is also possible to find reports that predict job losses of between 5 and 10 million jobs by 2020 [9]. According to a McKinsey report, technical work is more adaptable to automation. Jobs that require imagination, creativity, common sense, goal-setting and strategic thinking are harder to automate. Activities which involve managing and developing people for example have currently, according to McKinsey, only 9% automation potential, jobs that apply expertise to decision making, planning or creativity 18% [10]. The following figure illustrates the automation potential of jobs according to specific work activities:

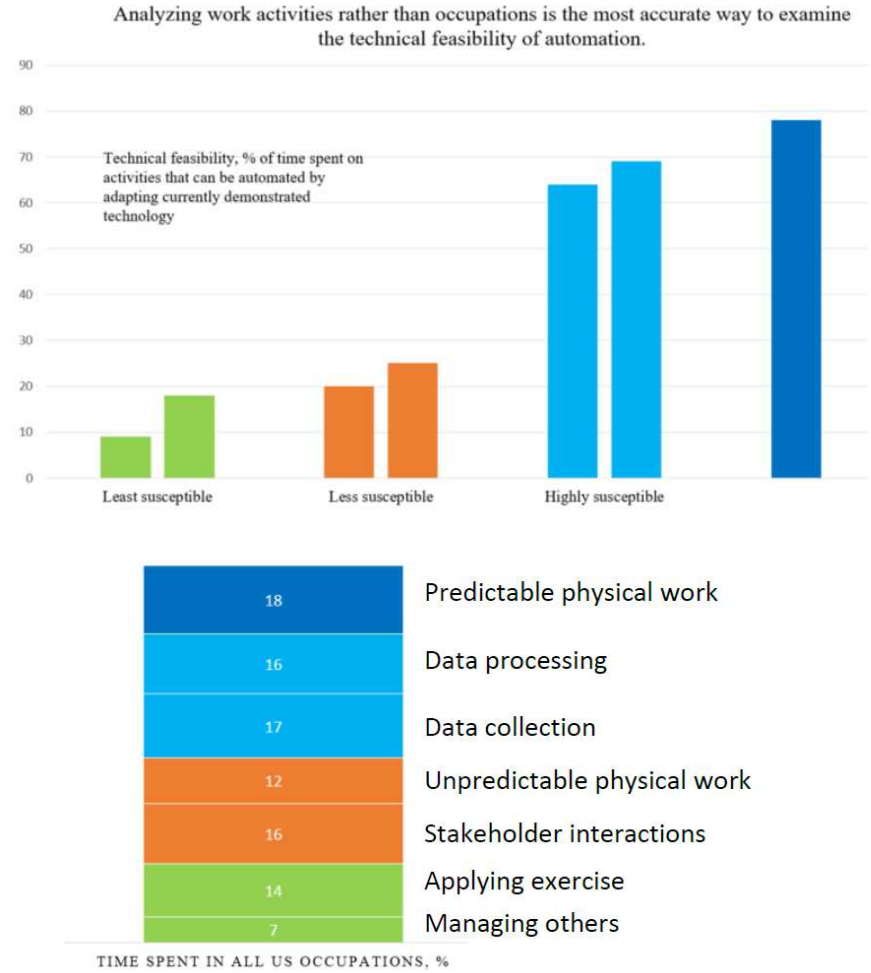


Figure 1. Work activities and automation potential [10]

There is much less research regarding the area of service robots and their effect on the labor market. A trend towards robots overtaking non-standardized tasks in this field of research can be noticed, due to technological advances. In addition, the prices for service robots are much lower than for industrial robots, which reduce the costs. Robots are not yet, but could become soon, a part of the value creation process in the service sector, even though this does not necessarily mean that they substitute human labor. They could also complement employees and make them more productive [11]. In a hospital for example, robots could

overtake transportation and logistics tasks and therefore the human workforce has more time available to spend with patients.

Maybe we do not necessarily have to be afraid to lose our jobs, due to an ongoing automation. Nevertheless, we will see a shift on the labor market and maybe not every employee is suitable for a job which requires creativity or leadership competence. The shift towards more service and social occupations will also mean that it could be necessary to re-educate employees and to change the perception and status of these jobs in our society. On the other hand, this could be a solution to labor shortage in these fields as well as an answer on how to deal with demographic change and an overageing society. Not only will there be more humans available to care for elderly people, less physically demanding jobs could also mean a chance for elderly people to be available on the labor market longer. Therefore, a structured and planned automation of work seems to be the reasonable solution to ensure, that these positive effects can be exploited without risking a de-stabilization of the labor market.

3. Military Robotics

New emerging technologies do not only change the labor market but have also the potential to change military and defense industry as well. Artificial intelligence and autonomous aerial and ground robots open up new ways of warfare. This also ethical raises questions like [12]

- Will autonomous robots be able to follow established guidelines of the Laws of War and Rules of Engagement, as specified in the Geneva Conventions?
- Will robots know the difference between military and civilian personnel?
- Will they recognize a wounded soldier and refrain from shooting?
- The ethical debate on military robotics is multi-layered. On the one hand, using robots in the battlefield could mean to preserve the lives of soldiers, who could have been harmed or killed otherwise.

The United Nations on the other hand urge to ban so-called “lethal autonomous weapons systems” before they are even developed [13]. Because of the increasing autonomy of robots, the prospect that systems could be used more flexible and therefore encounter influences which were not anticipated, as well as the complexity of technology, operational morality could not be sufficient for military robotics. Operational morality means that the actions of a robot are

completely in the hands of their designers or users and therefore the robot itself does not have to evaluate its actions regarding ethical concerns [12].

The effects of an unregulated development and use of these technologies can be illustrated using game theory. It could result in an AI arms race between superpowers like the US, Russia, the EU and China. Therefore, it seems necessary to prevent the development of such technologies before they exist, if this is not already too late by establishing regulations and not to found research in this field of expertise. Using robotics and artificial intelligence in military fields could mean advantages for the country or region regarding for example reaction time. This could result in a future where it is rational to ban human control almost entirely in military scenarios or applications [14].

The situation can be described by a social prisoner's dilemma scenario. If we look at two parties, e.g. the US and Russia, it is best for every single party to invest in military robotics and artificial intelligence. Nevertheless, the overall welfare is higher, if no one invests in these technologies [15]. The problem in real world scenarios is, that there is imperfect information and therefore it is difficult or even impossible to determine if the other party is collaborating, before it is to late. Therefore, it is "safer" for every single party to invest in these technologies, just in case the other party does so too. The following payoff matrix illustrates this by showing an example for possible outcomes of the individual country as well as overall outcomes, resulting from the different decisions:

Table 1. Example for a payoff matrix - prisoner's dilemma.

	County B Not investing		Country B Investing	
Country A Not investing	A: -2	B: -2	A: -6	B: -1
	Overall: -4		Overall: -7	
Country A Investing	A: -1	B: -6	A: -5	B: -5
	Overall: -7		Overall: -10	

Regulation, e.g. in the form of a code of conduct could help to eliminate information asymmetries and therefore make it more profitable for all involved parties not to invest in the development and use of artificial intelligence and military robotics in these fields. Nevertheless, in order to make the mechanism work, commitment of all parties as well as monitoring are crucial factors to build trust in these standards or guidelines [5].

On the other hand, not stopping these developments could result in a so-called "race to the bottom" regarding safety standards, as already illustrated in the case of military armament and private possession of weapons [14], [16].

4. Regional Effects

An ideal scenario would be to establish worldwide binding laws that will cover all critical questions and adapt to new developments without any time delay. But of course, this cannot be realized in practice. Therefore, Robotic Governance strives to provide and establish an ethical and moral framework for the new fields of robotics and artificial intelligence and therefore fill the gap that is necessarily emerging until such values will develop in society. This should empower individuals to make moral decisions as well as stakeholders and the society to identify parties that are not complying to these rules.

Nevertheless, as the Robot Manifesto is not binding, it is not realistic to expect it to be established worldwide at the same time. But if, in the first step, countries and regions like Europe, which is often a pioneer in the field of regulation and standards, or some important manufacturers and research institutes will voluntarily commit to these values, this can lead to so-called California Effect [17], known from the environmental sector, and the adoption of the code of conduct in other regions and companies. Here, the influence of the regions or stakeholders complying are important factors as well as, of course, the pressure from possible image damage in the public.

Another possible effect is the emergence of clusters. Again, two scenarios seem possible, looking at the field of environmental regulation, as a benchmark. On the one hand, it is reasonable to expect, that restrictions in research and development will lead to a competitive disadvantage, if other regions will invest in these technologies. The results would be robotic clusters in economic regions, which do not comply to these frameworks. Clusters are defined as a regional concentration of organizations, which are all connected by the same industry or field of activity. Due to the geographical proximity externalities and a potential for competitive advantages arise [18]. The so-called pollution haven hypothesis implies that companies from countries with a strict regulation will shift their production facilities abroad, in order to avoid these requirements [19]. This raises the question, if Robotic Governance will lead to robotic or technology havens. But here, the soft-regulation or self-regulation has advantages compared to a legal solution. The companies and organizations comply to these rules either because they are convinced that it is the right thing to do, e.g. because they were involved

in the process of establishing these guidelines, or they expect a competitive advantage from complying or at least a competitive disadvantage from not participating because of strong market pressure.

On the contrary, the Porter hypothesis suggests that regulation can also foster innovation - contrary to common neoclassical theories, which are based on the assumption, that a normal, profit orientated company will already have exhausted all available possibilities to increase efficiency and that additional regulation therefore would only involve additional costs [20]. Porter and van der Linde (1995) believe, that regulation can make companies more productive and encourage investments and technical progress in new directions. In addition, responsible organizations could gain a competitive advantage [21].

One might argue that especially in this context a variety of regulations could be beneficial for single states and regions. When Brandeis (1932) described the United States as laboratories of democracy, he acknowledged the benefits of allowing differing regulations, which can be adapted to the different needs and wishes of the population and therefore allow a greater freedom of choice [22]. Nevertheless, since 1932, Globalization changed a lot. The population is not always the customer base of the economy in a country or region. The effect would be, most likely, that some regions in developed countries offer sustainable products for a niche market that is very informed and could even gather a competitive advantage in this market. On the other hand, in other countries, where regulation is less strict, companies could produce e.g. technologies which do not meet certain ethical or moral standards and therefore be cheaper or more attractive for certain markets.

5. Conclusion

Of course, this is only an overview of some of the potential economic effects of technological advances in the fields of robotics and artificial intelligence. A complete analysis is almost impossible. Nevertheless, the examples illustrate why Robotic Governance and a Robot Manifesto – a Code of Conduct for the robotics community – might be more efficient than legislation, from an economic perspective, in order to solve the ethical and moral questions currently arising from the development of new technologies like artificial intelligence and autonomous machines.

The authors belief that, due to the fast pace of these developments, we have to create a uniform set of values regarding the development as well as use of robot and artificial intelligence, not to be overrun by the rapid technological progress in this fields, like this was the case with digitalization and the internet. We should

enable individuals to make moral decisions in particular cases, based on these accepted and trusted values and guidelines, instead of trying to regulate every single, ethically questionable case that might arise during the development of these new technologies. Especially because it is very hard to predict future innovations and technological breakthroughs as well as their effects in advance. In addition, if the technologies are developed, there is no way to undo them again.

The goal of the robotics community should be to develop technologies that will enhance the quality of life and serve humans, instead of posing a potential threat. Robotics and automation technologies have the potential to solve many of the problems like labor shortage and an overaging society, the economies of industrialized countries are currently facing. Robots could take on jobs that are dangerous or unpleasant for humans. Nevertheless, we should shape our future actively, instead of reacting to changes and trying to solve issues, after they emerged.

References

1. M. Arntz, T. Gregory, and U. Zierahn, *Digitalisierung und die Zukunft der Arbeit: Makroökonomische Auswirkungen auf Beschäftigung, Arbeitslosigkeit und Löhne von morgen*, ZEW (Zentrum für Europäische Wirtschaftsförderung GmbH), Germany, 04/04/2018 (2018).
2. G. H., Brundtland*, Our common future*, Report of the World Commission on environment and development, United Nations (1987).
3. I. Asimov, *I Robot*, New York: Doubleday & Company (1950).
4. D. Boesl, and M. Bode, *Generation 'R': Why our grandchildren will grow up as the first Generation of "Robotic Natives",* Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), IEEE, 417-420 (2016).
5. D. Boesl, and M. Bode, *Roboethics and Robotic Governance – A Literature Review and Research Agenda*, In: A. Ollero et al. (eds.), ROBOT 2017: Third Iberian Robotics Conference, Advances in Intelligent Systems and Computing 693 (2018).
6. A. Lyyra, *The idea of robots as independent machines is science fiction*, London School of Economics (2015).
7. J. J. Hughes, *What Is the Job Creation Potential of New Technologies?*, In: Surviving the Machine Age, 131-145 (2017).
8. M. Arntz, T. Gregory, and U. Zierahn, *The Risk of Automation for Jobs in OECD Countrie: a comparative analysis*, OECD Social, Employment and Migration Working Papers, **No. 189**, OECD Publishing (2016).
9. J. Pistrui, *The Future of Human Work Is Imagination, Creativity, and Strategy*, Harvard Business Review, 01/18/2018 (2018).

10. M. Chui, J. Manyika, and M. Miremadi, *Where machines could replace humans—and where they can't (yet)*, In: McKinsey Quarterly, 07/2016 (2016).
11. M. Decker, M. Fischer, and I. Ott, *Service robotics and human labor: A first technology assessment of substitution and cooperation*, In: Robotics and Autonomous Systems, **87** (2017).
12. P. Lin, G. Bekey, and K. Abney, *Autonomous military robotics: Risk,*
1. *ethics, and design*, California Polytechnic State University San Luis Obispo (2008).
13. O. Bowcott, *UN urged to ban 'killer robots' before they can be developed*, In: the Guardian, 04/09/2015 (2015).
14. T. Metzinger, *Towards a Global Artificial Intelligence Charter*, STOA - Science and Technology Options Assessment (2018).
15. P. D. Straffin, *The Mathematics of Tucker – A Sampler*, In: The Two-Year College Mathematics Journal, **14 (3),** 228–232 (1983).
16. J. Heath, and A. Potter, *The Rebel Sell*, Harper Collins, Toronto (2005).
17. D. Vogel, *Trading Up: Consumer and Environmental regulation in a global economy*, Harvard University Press (1995).
18. M. E. Porter, *Location, Competition, and Economic Development: Local Clusters in a Global Economy*, In: Economic Development Quarterly, **14(1)**, 15–34 (2000).
19. A. Levinson, and M. S. Taylor, *Unmasking the Pollution Haven Effect*, International Economic Review, **49 (1)**, 223–54 (2008).
20. S. Ambec, M. A. Cohen, S. Elgie, and P. Lanoie, *The Porter hypothesis at 20: can environmental regulation enhance innovation and competitiveness?* Review of environmental economics and policy**, 7(1)**, 2-22 (2013).
21. M. E. Porter, and C. van der Linde, *Toward a New Conception of the Environment-Competitiveness Relationship*, 99–101 (1995).
22. New State Ice Co. v. Liebmann, 285 U.S. 262 (1932).

ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.021

Demystifying "Value Alignment": Formally Linking Axiology to Ethical Principles in a Deontic Cognitive Calculus

Selmer Bringsjord, Naveen Sundar G and Atriya Sen

1. Introduction

A lot of people in and around AI who are concerned about immoral and/or destructive and/or unsafe AIs are running around calling for "value alignment" in such machines.* Our sense is that most of the time such calls are vapid, since those issuing the calls don't really know what they're calling for, and since the phrase is generally nowhere to be found in ethics itself. Herein we carry out formal work that puts some rigorous flesh on the calls in question. This work, quite limited in scope for now, forges a formal and computational link between axiology (the theory of value) and ethical principles (propositions that express obligations, prohibitions, etc.).

$\mathscr{CC}_{\mathcal{D}}$ is the space of deontic cognitive calculi. Hitherto, in particular calculi in this space that have been specified, and from there in some cases implemented, there is an absence of principled axiology. A case in point is $\mathcal{DCEC}^*$; this calculus, presented and used in (Govindarajulu & Bringsjord 2017*a*), specifies and implements what may so far be the most expressive, nuanced ethical principle to be AI-ified† — yet there is no principled axiology in this calculus, and hence none reported in the paper in question. Instead, an intuitive notion of positive and negative value, based on elementary arithmetic, and consistent with at least naïve forms of consequentialist ethical theories, is given and employed. In the present short abstract, we encapsulate our formally forging a link between Chisholm's

*E.g. see https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/.
†The Doctrine of Double-Effect.

(1975) intrinsic-value axiology,[‡] and ethical principles, in order to introduce the road to doing this in a rich way, subsequently.

2. Chisholm's Intrinsic-Value Axiology, Absorbed

Chisholm (1975) gives a theory of *intrinsic* value expressed in a quantified propositional logic that takes as primitive a binary relation *Intrinsically Preferable*; this allows him to e.g. write $P(p, q)$, where here P is the primitive relation, and p and q are of course propositional variables. To absorb Chisholm's system into a deontic cognitive calculus, we begin by immediately recasting Chisholm's theory in quantified multi-modal logic, in which all his first-order relations are modal operators, and his propositional variables are variables ranging over unrestricted formulae (which may themselves have modal operators and quantifiers within them). E.g., P becomes the binary modal operator $\mathbf{Pref}$.[§] We cast Chisholm's six definitions as axioms, and translate his five axioms as described immediately above. This yields 11 axioms, specified as follows:

A1 $\forall \phi, \psi\ [\mathbf{Same}(\phi, \psi) \leftrightarrow \neg\mathbf{Pref}(\phi, \psi) \wedge \neg\mathbf{Pref}(\psi, \phi)]$
A2 $\forall \phi\ [\mathbf{Indiff}(\phi) \leftrightarrow \neg\mathbf{Pref}(\phi, \neg\phi) \wedge \neg\mathbf{Pref}(\neg\phi, \phi)]$
A3 $\forall \phi\ [\mathbf{Neutral}(\phi) \leftrightarrow \exists \psi\ (\mathbf{Indiff}(\psi) \wedge \mathbf{Same}(\phi, \psi)]$
A4 $\forall \phi\ [\mathbf{Good}(\phi) \leftrightarrow \exists \psi\ (\mathbf{Indiff}(\psi) \wedge \mathbf{Pref}(\phi, \psi)]$
A5 $\forall \phi\ [\mathbf{Bad}(\phi) \leftrightarrow \exists \psi\ (\mathbf{Indiff}(\psi) \wedge \mathbf{Pref}(\psi, \phi)]$
A6 $\forall \phi, \psi\ [\mathbf{ALAG}(\phi, \psi) \leftrightarrow \neg\mathbf{Pref}(\psi, \phi)]$
A7 $\forall \phi, \psi\ [\mathbf{Pref}(\phi, \psi) \rightarrow \neg\mathbf{Pref}(\psi, \phi)]$
A8 $\forall \phi, \psi, \gamma\ [(\mathbf{ALAG}(\psi, \phi) \wedge \mathbf{ALAG}(\gamma, \psi)) \rightarrow \mathbf{ALAG}(\gamma, \phi)]$
A9 $\forall \phi, \psi\ [(\mathbf{Indiff}(\phi) \wedge \mathbf{Indiff}(\psi)) \rightarrow \mathbf{Same}(\phi, \psi)]$
A10 $\forall \phi\ [(\mathbf{Good}(\phi) \wedge \mathbf{Bad}(\neg\phi)) \rightarrow \mathbf{Pref}(\phi, \neg\phi)]$
A11 $\forall \phi, \psi\ [\neg(\mathbf{Pref}(\phi, \phi \vee \psi) \wedge \mathbf{Pref}(\psi, \phi \vee \psi)) \wedge \neg(\mathbf{Pref}(\phi \vee \psi, \phi) \wedge \mathbf{Pref}(\phi \vee \psi \vee \psi))]$

3. Some Axiological Theorems, Machine-Discovered/Verified

In order to obtain some object-level theorems from A1–A11, we of course must have a proof theory to anchor matters; and in order for these theorems to be automatically obtained we shall of course need this theory to

[‡]The earlier version of which is (Chisholm & Sosa 1966).
[§]Later, it will become necessary to move beyond Chisholm by allowing this operator to take into account agents α, and thus it will become ternary: $\mathbf{Pref}(\phi, \psi, \alpha)$. The other operators in the modalized axiology of Chisholm would of course need to include a placeholder for agents.

be implemented. We use a simple proof theory, $\mathcal{R}_{\mathcal{A}}$, for now. Our first ingredients are inference schemata needed for quantificational reasoning over the 11 axioms. We thus allow the standard natural-deduction schemata allowing introduction and elimination of $\exists$ and $\forall$ in A1–A11; and of course we allow the remaining natural-deduction schemata for first-order logic: *modus ponens*, indirect proof, and so on, where the wffs allowed in these schemata may contain operators. In this mere abstract, we don't discuss any of the theorems that are now reachable, nor do we show that these theorems can be automatically proved by ShadowProver (Govindarajulu & Bringsjord 2017*b*, Govindarajulu, Bringsjord, Ghosh & Peveler 2017).

We in addition invoke the inference schemata (resp., as S_g and S_b)

$$\frac{\vdash_{\mathcal{R}_{\mathcal{A}}} \phi}{\mathbf{Good}(\phi)}$$

and

$$\frac{\vdash_{\mathcal{R}_{\mathcal{A}}} \neg\phi}{\mathbf{Bad}(\phi)}$$

4. Bridging to Ethical Principles

So far there is no connection between value and traditional ethical categories, such as the *obligatory* and *forbidden*. In the full paper, we introduce "bridging" principles that take us from value to not only these two categories, but to the complete spectrum of ethical categories in (Bringsjord 2015). Here, put informally, are two of the principles we formalize and explore:

P1/P2 Where ϕ is any good (bad) state-of-affairs, it ought to be (is forbidden) that ϕ.

5. Some "Value-Alginment" Theorems, Machine-Discovered/Verified

In the full paper, we explore the automated proving of the theorems in §4 by way of ShadowProver.

6. On Deriving an 'Ought' From an 'Is'

Hume famously maintained in his *Treatise of Human Nature* that an ought cannot be dervied from an is. Yet it appears that, one, those calling for

"value alignment" are calling for what Hume declared impossible, and that, two, we have nonetheless accomplished the very thing, for we have:

> **Theorem**: It ought to be that $\phi \rightarrow \phi$.
>
> **Proof**: Trivial: $\phi \rightarrow \phi$ is a theorem, and hence by S is **Good**, and thus by S_g is obligatory. **QED**

7. Conclusion and Next Steps

We have sought to explicitly and formally forge a connection between axiology and deontic concepts and propositions, in order to rationalize calls for "value alignment" in AI. Have we succeeded? At this point, confessedly, the most that could be said in our favor is that we have put on the table an encapsulation of a candidate for such forging. What additional steps are necessary?

Some are rather obvious. Goodness of states-of-affairs, and badness of them as well, would seem to fall into continua; yet what we have adapted from Chisholm and Sosa allows for no gradations in value. That goodness and badness comes in degrees appears to be the case even if attention is restricted to states-of-affairs that are intrinsically good (bad). For instance, knowledge (at least of "weighty" things, say the stunning truths about the physical world in relativity theory) on the part of a person would seem to have intrinsic value, but the selfless love of one person for another would seem to be something of even greater intrinsic value. Yet at this point, again, our axiology admits of no gradations in goodness and badness. We know that we need a much more fine-grained axiology.

Our suspicion is that goodness and badness should be divided between what has intrinsic value/disvalue, and what has value instrumentally. Instrumental value would be parasitic on intrinsic value. More specifically, we are inclined to think that both the instrumental and the intrinsic is graded.

One final remark: Even at this early phase in the forging of a formal connection between value and ethical principles based on deontic operators, it seems patently clear that because actual and concrete values in humanity differ greatly between group, nation, culture, religion, and so on, any notion that a given AI or class of AIs can be aligned with *the* set of values in humanity isn't only false, but preposterous. For many Christians, for example, the greatest intrinsic good achievable by human persons is direct and everlasting commuion with the divine person: God. For many others (e.g. (Thagard 2012)), this God doesn't exist, and the greatest goods are achieved by living, playing, and working in the present worldm in which our lives are short and end forever upon earthly death. In the context formed

by the brute fact that values among human persons on our planet vary greatly, and are, together, deductively inconsistent, our focus on formality is, we submit, prudent. Once the formal work has advanced sufficiently, presumably the alignment of AIs with values can be undertaken relative to a concretely instantiated axiology, from which ethical principles flow. For a given class of AIs, then, their behavior would be regulated by ethical principles that flow from a particular instantiation of the axiology.

References

Bringsjord, S. (2015), A 21st-Century Ethical Hierarchy for Humans and Robots: $\mathcal{EH}$, *in* I. Ferreira, J. Sequeira, M. Tokhi, E. Kadar & G. Virk, eds, 'A World With Robots: International Conference on Robot Ethics (ICRE 2015)', Springer, Berlin, Germany, pp. 47–61. This paper was published in the compilation of ICRE 2015 papers, distributed at the location of ICRE 2015, where the paper was presented: Lisbon, Portugal. The URL given here goes to the preprint of the paper, which is shorter than the full Springer version. **URL:** *http://kryten.mm.rpi.edu/SBringsjord_ethical_hierarchy_0909152200NY.pdf*

Chisholm, R. (1975), 'The Intrinsic Value in Disjunctive States of Affairs', *Noûs* **9**, 295–308.

Chisholm, R. & Sosa, E. (1966), 'On the Logic of "Intrinsically Better"', *American Philosophical Quarterly* **3**, 244–249.

Govindarajulu, N. & Bringsjord, S. (2017*a*), On Automating the Doctrine of Double Effect, *in* C. Sierra, ed., 'Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)', International Joint Conferences on Artificial Intelligence, pp. 4722–4730. **URL:** *https://doi.org/10.24963/ijcai.2017/658*

Govindarajulu, N. S. & Bringsjord, S. (2017*b*), On Automating the Doctrine of Double Effect, *in* C. Sierra, ed., 'Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17', Melbourne, Australia, pp. 4722–4730. Preprint available at this url: `https://arxiv.org/abs/1703.08922`. **URL:** *https://doi.org/10.24963/ijcai.2017/658*

Govindarajulu, N. S., Bringsjord, S., Ghosh, R. & Peveler, M. (2017), Beyond The Doctrine Of Double Effect: A Formal Model of True Self-Sacrifice. International Conference on Robot Ethics and Safety Standards.

Thagard, P. (2012), *The Brain and the Meaning of Life*, Princeton University Press, Princeton, NJ.

AUTHOR INDEX

ISBN: 978-1-9164490-1-5