ICRES2021
6th International Conference on Robot Ethics and Standards
New York, USA, 26-27 July 2021

Life-world for Artificial and Natural Systems

Editors
Selmer Bringsjord
Mohammad O. Tokhi
Maria Isabel A. Ferreira
Naveen S. Govindarajulu
Manuel F. Silva

LIFE-WORLD FOR ARTIFICIAL AND NATURAL SYSTEMS

LIFE-WORLD FOR ARTIFICIAL AND NATURAL SYSTEMS

ICRES 2021 Proceedings,
New York, USA, 26-27 July 2021

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

Manuel F. Silva
Porto Polytechnic, Portugal

Published by

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

Life-world for Artificial and Natural Systems
Proceedings of the Sixth International Conference on Robot Ethics and Standards

PREFACE

ICRES 2021 is the sixth edition of the International Conference series on Robot Ethics and Standards. The conference is organized by CLAWAR Association in collaboration with Rensselaer Polytechnic Institute (RPI), and held in New York, USA on a virtual platform during 26 – 27 July 2021.

ICRES 2021 brings new developments and new research findings in robot ethics and ethical issues of robotic and associated technologies. The topics covered include fundamentals and principles of robot ethics, social impact of robots, human factors, regulatory and safety issues.

The ICRES 2021 conference includes a total of five plenary lectures, and 15 regular and invited presentations. A special discussion panel session on Automation of Machine Ethics is also organised.

The editors would like to thank members of the International Scientific Committee and Local 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 2021 proceedings will be a valuable source of reference for research and development in the rapidly growing area of robotics and associated technologies in context of ethics and standardisation framework.

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

CONFERENCE ORGANISERS

CLAWAR Association
www.clawar.org

Rensselaer Polytechnic Institute in Troy, New York, USA
https://www.rpi.edu/

CONFERENCE COMMITTEES AND CHAIRS

Conference Chairs and Managers

Selmer Bringsjord (General Co-Chair)	– Rensselaer Polytechnic Institute, USA
Mohammad Osman Tokhi (General Co-Chair)	– London South Bank University, UK
Maria Isabel Aldinhas Ferreira (General Co-Chair)	– University of Lisbon, Portugal
Gurvinder S. Virk (Int. Advisory Committee)	– CLAWAR Association, UK
Abdullah Almeshal (Web-site)	– College of Technological Studies, Kuwait

International Scientific Committee

Naveen S. Govindarajulu (Co-Chair)	– Rensselaer Polytechnic Institute, USA
Manuel F. Silva (Co-Chair)	– ISEP & INESCTEC, Portugal
Ronald Arkin	– Georgia Institute of Technology, USA
Raja Chatila	– ISIR/UPMC-CNRS, France
Dimitris Chrysostomou	– Aalborg University, Denmark
Roeland de Bruin	– Utrecht University, The Netherlands
Sarah Fletcher	– Cranfield University, UK
Khaled Goher	– University of Lincoln, UK
Joseph Johnson	– Rensselaer Polytechnic Institute, USA
Endre E. Kadar	– University of Portsmouth, UK
Aman Kaur	– London South Bank University, UK
Philip Lance	– PA Consulting, UK
João Sequeira	– University of Lisbon, Portugal
Murray Sinclair	– Loughborough University, UK
Alan Winfield	– University of the West of England, UK
Karolina Zawieska	– Industrial Research Institute for Automation and Measurement, Poland

Organising Committee

Naveen S. Govindarajulu (Chair)	– Rensselaer Polytechnic Institute, USA
Abdullah Almeshal	– College of Technological Studies, Kuwait
Dimitris Chrysostomou	– Aalborg University, Denmark
Mike Giancola	– Rensselaer Polytechnic Institute, USA
Khaled Goher	– University of Lincoln, UK
Aman Kaur	– London South Bank University, UK

TABLE OF CONTENTS

Section–3: Special session presentations on "Data Analytics, a tool for development: The technical, ethical and legal complexities"

Section–4: Special session presentations on "The psychology of the life-world for autonomous agents"

Section–5: Regular presentations

ICRES 2021: Sixth International Conference on Robot
Ethics and Standards, New York, USA, 26-27 July 2021.

SECTION-1: PLENARY PRESENTATIONS

Are we ready to expect people to work with industrial robots?
Sarah Fletcher, Cranfield University, UK

Robots and AI in the real world: Applications and challenges
Amit Kumar Pandey, Hanson Robotics, Hong Kong

Never before in history, Robots, AI and IoT, all together have been so close to us, in our society. It is a revolution towards a new ecosystem of living, where AI is now the part of our lives and robots are catching up already. The intention is to facilitate a smarter, healthier, safer and happier life. Such artificial intelligent beings getting used in education, healthcare, retail, entertainment, art, science, and even to improve our understanding about ourselves, the human being. The talk will focus on some of such potential use cases, provide industrial and end users perspective, and discuss the scientific, technological and, social and ethical challenges we need to address as a community. The talk will open the floor by highlighting the multidisciplinary nature of the domain, and the need of a bigger collaborative ecosystem.

Robotics and AIOT: Technical and ethical challenges
Sule Yildirim Yayilgan, Norwegian University of Science and Technology, Norway

Attention mechanisms and self-awareness in intelligent systems
Arlindo Oliveira, Instituto Superior Técnico, Portugal

AI ethics standards and regulation for a trustworthy AI ecosystem
Ansgar Koene, University of Nottingham, UK

SECTION -2: INVITED SESSION PRESENTATIONS "HUMAN/ROBOT COOPERATION IN INDUSTRIAL SETTINGS – THE WAY FORWARD"

Human Robot Collaborative Applications – Evolution of challenges vs technological solutions
George Michalos, Patras University, Greece

Human - robot collaboration using visual cues for communication
Iveta Eimontaite, Cranfield University, UK

Development of adaptive and collaborative human-robot systems exploiting context-based information
Angelo Marguglio, Engineering Ingegneria Informatica SPA, Italy

Safety controllers in human-robot collaboration: Verified synthesis
Mario Gleirscher, University of Bremen. Germany

Active robot assistance with mutual understanding by predictability
Kevin Haninger, Fraunhofer Institute. Germany

On human condition: The status of work
Maria Isabel Aldinhas Ferreira, University of Lisbon. Portugal

Risk assessment in HRC in industry
Elena Dominguez, Pilz, USA

Robots and the Workplace: The contribution of technology assessment to their impact on work and employment
Tiago Carvalho, Colabor, Portugal

SECTION -3: INVITED SESSION PRESENTATIONS "DATA ANALYTICS, A TOOL FOR DEVELOPMENT: THE TECHNICAL, ETHICAL AND LEGAL COMPLEXITIES"

Is Inferred Data Private?

Selmer Bringsjord & Naveen Sundar G., Renssalaer Polytechnic Institute

A certain rather small declarative database Y holds plenty of personal information about you. What information? Someone you happened to meet briefly by chance at a restaurant built Y after this meeting. The man you met was a pleasant French gentleman, a most polite and refined Monsieur Dupin, of Paris. You were with your spouse, and the two of you were invited to have a seat at the bar for an aperitif while your table was cleared and configured for the start of a gourmet dinner in Porto, Portugal, overlooking the great river itself. Dupin happened to be at the bar already, and you sat down next to him, with your spouse at your side; these were the last two available seats. You volunteered your full name to Dupin, and he knew immediately from that that while your residence at present isn't necessarily New York, you grew up there in large measure. When your spouse volunteered her name, only her first, he knew that the two of you both grew up within a relatively short distance of each other. Dupin also later installed in Y that your spouse took your last name upon marrying you. Dupin's Y also contains your age and that of your spouse as well, plus or minus (as noted in Y) five years. Dupin observed that both your spouse's purse and her shoes were designer brands, and that your watch was an Omega Speedmaster. Dupin generously insisted upon buying the both of you your apertifs, and asked what your preference would be between a glass of A versus a glass of B, both from — as he put it — his "backyard" back in France. He started an internal timer running as he awaited your reply. Almost instantaneously and without missing a beat, you replied that this offer was very kind, and that you both would have B — which Dupin knew to be markedly less expensive than A; still quite dear, but much, much less.

Dupin was later able to in fact add to Y an astonishing amount of data that he inferred from what as noted above he learned during your brief time together at the bar. (Do you see how?)

Now, is what Y holds private data regarding you and your spouse? Is Dupin's assembling Y a violation of your privacy? I shall answer these questions both firmly in the negative, and in defense of these answers employ a sorities-style argument, one that begins from a particular Dupin-inferred fact in Y, namely that you know the aperitif (wonderful) options on the menu from north of the great river are "fizzy" because of added carbonation.

Bias as limitation of modelling the world

Luís Mateus Rocha, State University of New York at Binghamton, USA

Processing AI-data - mind the web: European (proposed) regulations

Roeland de Bruin, Utrecht University, The Netherlands

It is evident that data analytics forms the cornerstone of self-learning and other AI-related algorithms. Large scale data processing requires as little problems as possible regarding input, processing and output. This has, among many other things, recently formally been underscored by the regulatory institutions of the European Union, who drafted a proposal for harmonized rules on Artificial Intelligence. In this Special Session, I will focus on the proposed rules to the extent they regard the use and governance of data, and the intersection with the ever more elaborated EU rules regarding personal data protection.

The Proposed AI-regulation – a data perspective

In the proposed regulations, a normative framework is drafted which should on the one hand stimulate consumer trust and on the other hand should stimulate innovation by providing clear rules for innovators. The EC states that it is beneficial for trust when inter alia the fundamental right to privacy of citizens is duly observed by AI-innovators – in conformity with the data protection framework, including the General Data Protection Regulation (GDPR), alongside the other fundamental rights catalogue of the Charter of Fundamental Rights of the European Union. Other aspects that may negatively impact trust have to be avoided. It is proposed that harmful AI systems are prohibited, and high risk AI systems may only be deployed on the basis of a risk management system, in order to reduce consumer risks as much as possible throughout the lifetime of an AI system. It furthermore provides rules for the input data and the governance and management thereof. In that, due account must be also be taken of the General Data Protection Regulation. The proposed obligations for AI systems providers will be assessed during the Special Session, especially where it concerns personal data, and these can be related with the obligations for controllers and processors of personal data following from the GDPR.

The General Data Protection Regulation – revisited

In the past three years, ever more "open norms" of the GDPR are filled in by guidance of the European Data Protection Board (EDPB), and case law of the Court of Justice of the European Union (CJEU). The CJEU decision in the Schrems II case is particularly relevant, as it concerns the export of personal data from the EU to third countries, especially the United States. Significant constraints result from the annulment of the so called Privacy Shield, which will be assessed during the Special Session. In that, we will also relate to the interpretation of the Schrems II-decision by the EDPB. The question will be raised to what extent AI-related personal data processing would still be allowed when a processor (or controller) from the US is involved in the processing chain. As this certainly is not the only development that is relevant from an AI-perspective, we will also zoom in on inter alia the EDPB guidelines regarding "processing personal data in the context of connected vehicles and mobility related applications".

Towards ethical AI in mission critical applications

Muhannad Alomari, Rolls Royce, UK

Business to local government data sharing

Anthony Colclough, Urban Expert, Eurocities

Local governments can unlock enormous benefits for local people with access to the right data. This happens not only when cities use the data for improving public services, but also when they act as facilitators, making data available to civil society, small and medium enterprises and entrepreneurs to develop their own bottom-up ideas. However, local government often finds that it is not in the position to access or share data generated on its territory, even when such data is generated by companies publicly tendered to provide local services. How do cities navigate their own access to data, and making that data available to others? And what can some best practice examples from cities tell us about the way forward? Through the presentation of a recent paper developed in close consultation with the cities of Eurocities' Knowledge Society Forum, and success stories from southern San Sebastian and northern Turku, this intervention will elucidate a cities' perspective on contemporary data governance.

The FBPML Open-source best practices

Jeroen Franse, Foundation for Best Practices in Machine Learning

The acknowledged operational, ethical, legal and governance risks have generated a need for a clear and thoughtful repository of best practices on how to responsibly govern, manage and implement Machine Learning ("responsible ML").

The Foundation for Best Practices in Machine Learning (non-profit) seeks to promote responsible ML through creating an open-sourced, freely accessible repository of best practices

and associated guides. Its model and organisational guides look at both the technical and institutional requirements needed to promote responsible ML. Both blueprints touch on subjects such as “Fairness & Non-Discrimination”, “Representativeness & Specification”, “Product Traceability”, “Explainability” amongst other topics. Where the organisational guide relates to organisation-wide process and responsibilities (f.e. the necessity of setting proper product definitions and risk portfolios); the model guide details issues ranging from cost function specification & optimisation to selection function characterization, from disparate impact metrics to local explanations and counterfactuals. It also addresses issues concerning thorough product management.

These guidelines have been developed principally by senior ML engineers, data scientists, data science managers, and legal professionals for ML engineers, data scientists, data science managers, compliance professionals, legal practitioners, and, more broadly, management. The Foundation’s philosophy is that (a) context is key, and (b) responsible ML starts with prudent MLOps and product management.

Robots, Standards, Ethics and privacy: the coexistence in a legal perspective through the model DAPPREMO

Nicola Fabiano, International Institute of Informatics and Systemics (IIIS), USA

Still, nowadays, some people consider Robots or Robotics and technical standards as domains typically falling exclusively in technicians' competence. The common experience teaches us that we can face those domains even whether we have to deal with the legal ones. Multidisciplinarity does not mean debasing professionalisms; people from a technical background will continue to work with excellent achievements like those from the legal domains.

A radical change of mentality is required.

Recently the European Union has approved a proposal of regulation on Artificial Intelligence. But what is Artificial Intelligence?

To get a satisfactory legal framework capable of dealing with this multifaceted phenomenon, a multidisciplinary approach is required. The hybridisation of competencies does not mean relinquishing the sphere proper to the jurist, but it allows to get a precise view on the nature and composition of what is always a multilayred context.

Robotics and Artificial Intelligence domains represent the leading technological research area of our future. The challenge is to welcome and encourage any innovation by balancing with Ethics, Data Protection and Privacy.

In this talk we will introduce DAPPREMO (acronym for Data Protection and Privacy Relationships Model) which has proved to be helpful providing a broader vision of the entire reality around a single case. In that way, we can have clear in advance what domains we should deal with them. The following step will establish the proper procedure or process to work on the entire context with the right approach and carry out the maximum result. It is not very easy to join robots, technical standards, ethics and privacy (or data protection) laws. We think they can coexist by adopting the model named DAPPREMO (acronym of Data Protection and Privacy Relationships Model).

SECTION -4: INVITED SESSION PRESENTATIONS "THE PSYCHOLOGY OF THE LIFE-WORLD FOR AUTONOMOUS AGENTS"

According to Encyclopedia Britannica Life-world, German Lebenswelt, is defined the following way: "the world as immediately or directly experienced in the subjectivity of everyday life, as sharply distinguished from the objective "worlds" of the sciences, which employ the methods of the mathematical sciences of nature; although these sciences originate in the life-world, they are not those of everyday life."

Husserl deserves credit for coining the term "life-world" and highlighting the limitations of standard scientific methods and assumptions in handling the problems of life-world. Robotic researchers should be keenly aware of these problems whilst designing autonomous agents that fit into the life-world. In particular, it is imperative to take the psychology of the life world seriously in safe and ethical designs of robots. The present symposium provides an overview of some of the key aspects of the challenges robotics researchers have to face.

Robustness and variability of the behaviour of autonomous agents
Endre E Kadar and Danielle Foxley, University of Portsmouth, UK

ICRES 2021: 6th International Conference on Robot
Ethics and Standards, New York, USA, 26-27 July 2021.

WORKING TOGETHER WITH ROBOTS TO IMPROVE LEARNING

TIMOTHY GIFFORD
Movia Robotics, Inc. 72 Prospect Pl
Bristol, CT 06010, USA
E-mail: tgifford@moviarobotics.com
www.moviarobotics.com

Collaborative robotics provides a unique opportunity for educating children. This intervention technique has raised concerns over the ethical implications of teaching children social skills with a mechanical device. Will the children prefer to interact with a robot and turn away from human interaction? Will robot-based therapies replace human therapists, taking the humanity out of the child's education? Providing successful training interventions to children with ASD is very challenging. Children with ASD have difficulty maintaining engagement and attention. They often do not like social interaction. Robot-Assisted Instruction (RAI) overcomes some of these challenges. Children find robots engaging and often treat them as a social entity. Social interactions are perhaps the most complex processes in the life-world. Reading gestures and facial expressions can be overwhelming by presenting multiple social cues simultaneously. Children with ASD are struggling to cope with this information overload. The robots are not able to present the complexity of these social cues, which provides an opportunity to teach social skills one aspect at a time. The robot and associated devices act in concert with the facilitator enabling them to work as a team as they teach the child. In RAI, the robot leads the children through training interventions giving the children experience in activities related to social emotional learning, learning readiness, activities for daily living and academics. The robot provides an opportunity for the child to practice social interactions in a safe and predictable environment. The systems are semiautonomous giving the facilitator a powerful tool to provide educational instruction while maintaining control over the system to insure appropriate application. The system can dynamically change its role to meet the child's level of engagement and interaction. This combined with the participation of the facilitator both through control and adjustment of the system and directly with the child through prompting ensures that the intervention is safe and comfortable while being efficacious for the child. The robot and human facilitator team provide training experiences that improve the outcomes for children by providing a safe and comfortable practice environment. The skills acquired are generalized and used by the child in situations where the robot is not present.

1. Background

1.1. *Autism Spectrum Disorder (ASD)*

ASD affects children in many different ways. Our approach pays particular attention to issues involving interactions across multiple modalities including interpersonal coordination through movement and gestures. Children with ASD who have imitation impairments at a young age also present with language delays in the preschool years (Stone & Yoder, 2001). Imitation deficits in young and older children with ASD correlate with their other social skills such as joint attention (i.e., ability to coordinate attention between people and objects) and their understanding of others' intentions (Mundy, Sigman, & Kasari, 1990; Baron-Cohen & Swettenham, 1997; Sigman & Ruskin, 1999; Charman et al., 2003). Imitation training, such as reciprocal imitation and visually cued imitation, improves the social communication skills of children with ASD (Ingersoll & Gergans, 2007; Ingersoll, Lewis, & Kroman, 2007; Ganz et al., 2008). Children with high functioning ASD showed fewer correct responses during gestures following imitation, gestures to command, and gestures during tool use (Mostofsky et al, 2006). Young and older children with low and high functioning ASD have impaired fine and gross motor coordination including basic motor skills such as locomotion and upper limb tasks as well as static and dynamic balance tasks (Ghaziuddin, et al., 1994; Henderson & Sugden, 1992).

1.2. *Motor Performance and Joint Attention*

Findings indicate that enhancing the motor performance of children with ASD may facilitate their poor social communication skills (Sutera et al., 2007; Brian et al., 2008; Gernsbacher et al., 2008). Joint attention (JA) is the ability to focus one's attention to that of a social partner (Mundy & Sigman, 2006). Children with ASD have deficits in appropriately responding to JA. Studies suggest that spontaneously initiating JA is significantly impaired in children with ASD. Four-year old children with ASD improved their response and initiation of joint attention behaviors following joint attention training (Whalen & Schreibman, 2003). Another study found that young children with ASD make significant gains in language development following JA based intervention as compared to an untrained control group (Kasari et al., 2008). Due to the JA and other skills deficits with children on the Autism spectrum and educator can have a very difficult time with the engagement of Autistic students in a typical classroom setting. These difficulties can be both academic and behavioral in nature. These difficulties are further enhanced by the anxiety a child with Autism can have being in a classroom because of the inability to handle the various situations and there needed responses that occur throughout the school day.

1.3. *Effects of Robot Interactions*

Research has shown that children with autism have a unique affinity towards robots. This is evidenced by their willingness to engage and interact with the robots socially. Several researchers have shown that children with ASD may demonstrate more engagement with robots than with humans (Robins, Daughtenhahn, & Dubowski, 2006); Bekele et al., 2013, and Kim et al., 2012). This has opened the opportunity to lead the children through productive learning activities. Further research confirmed the robustness of the engagement to robots seen in children with ASD (Toh, 2016). Beyond engagement there are many beneficial effects for the child when interacting with a robot. Multiple studies have shown an increase in compliance within participants after working with robots. (Bainbridge, 2008; Srinivasan et al., 2015). Research has also shown an increase cognitive learning gain (Leyzbeg, 2012). Also, children with ASD produce higher rates of joint attention that are comparable to typically developing children when interacting with robots (Kim et al., 2012; Pop et al.). Importantly there is evidence that these skills are generalized and present in the participants when the robots are not present. This was shown to be true for social skills where the children demonstrated generalization of social skills with people, including eye contact (Scassellati, 2018).

1.4. *Robot as Embodied Social Interactor*

Research demonstrates that children with ASD produce more vocalizations when engaging with robots than with other humans or a computer screen. This was shown in a study where students exhibited increased verbalization and socialization with an embodied robot versus a screen-based app and were more socially comfortable than with humans (Kim, 2013). Part of this effect might be due to the stimulation of mirror neurons when an embodied entity occupies the same space as the participant. (Gazzola, 2007). This increased brain activity leads to an increased engagement in motor skill activities and joint attention (Tapus, 2012). Robot based intervention can target joint attention behaviors during triadic interactions between the child, the tester or teacher, and the robot with the robot as the object of JA. Robot based interventions can be also used to facilitate complex motor coordination and postural control of children through imitation. Robots can be used to facilitate action imitation and interpersonal coordination. Research in

embodied cognition shows that joint coordination activities improve interpersonal coordination (Marsh et al, 2009) Research using robots with children in joint movement activities shows gains in interpersonal coordination as well as spontaneous appropriate verbalizations (Srinivasan et al., 2015; Kaur et al, 2013).

1.5. *Benefit as Assistive Technology*

Robot Assisted Instruction systems provide the basis for a deployable assistive technology system for working with students with ASD in the school, clinic or home environment. The ability of the robot to lead the child through training interventions leaving the specialist free to direct and observe the interactions is beneficial to the child and the specialist. The child finds the interactions more enjoyable and accessible with the potential for more time on task. Having the robot lead the activities gives the therapist a better opportunity for observation and to collect data while dynamically assessing the progress of the child. The objective nature of the robot interaction also removes some of the variability of delivery. Children with ASD maintain good engagement with RAI over long periods of time. Skills learned with the robot are generalized and repeated by the child when the robot is not present.

2. Method

2.1. *Robot, Child, and Facilitator Grouping*

The robot, child and facilitator form a synergistic group that interact with each other. The robot leads the child through educational activities including lessons and games. The facilitator interacts with the child and controls certain aspects of the robot's behavior through the controller. These 3 players interact as a group. The robot acts in a semi-autonomous way providing interactions in a linear fashion with dynamic modifications to delivery and complexity. The child interacts through speech, movements and by pressing graphical icons on a tablet. The tablet provides a way for the child to input responses that is unambiguous. It enables the robot to respond without the necessity for speech recognition. This is important as speech recognition software is prone to mistakes especially with children and those with speech impediments. The speech and movement interactions are interpreted by the facilitator and then input into the system. The robot makes the appropriate response based on the inputs from each source.

2.2. *Roles and Context in Defining Interactions*

The robot is perceived as an animate social entity by the child. The robot and child interact within the context of a social interaction. The context changes throughout the session but it always remains consistent to what is appropriate at that time. The robot will take on different roles based on the needs of the moment. Sometimes the robot is a teacher and sometimes it is a playmate. The robot has the ability to change between roles by changing its mode of operation. Each mode is a state of equilibrium where the robot can proceed along a reduced set of action possibilities. These action possibilities fully describe what is appropriate for the robot in each role. Since the robot has only a few action possibilities within a specific role it is possible for the semiautonomous control system of the robot to make appropriate action choices. The robot will act proactively, leading the child through the activities. If the child changes her behavior to something that is not within the context of the activity the robot can switch to another role or state of operation. This new role can be to try to return the child to desired state of participating

in the lesson or the new role can follow the child to a new action state. This state meets the child at their current level. For example this new role could be one of playmate or of de-escalation. Here the robot will interact with the child proactively to bring the child to a more favorable state of learning readiness.

2.3. *Multimodal Interactions*

The interactions between the robot and child are multimodal in the form of speech, movement gestures, expressions and sounds. The tablet displays graphical representations of concepts that are being taught. The tablet also provides graphical buttons for the child to choose from when answering questions. These multiple forms of interaction provide many opportunities for engagement and rapport building between the child and robot. These coordinated activities also provide opportunities for shared experience between the child and facilitator. These experiences can be shard in the moment through joint attention bids and later as memories for storytelling and other pragmatic communication activities.

2.4. *Interaction Structure*

The interactions between the robot and child are structured to support specific types of engagements. The structure is maintained from session to session to give the child a predictable yet dynamic experience. This is important to provide a comfortable experience for the child while maintaining novelty. The robot engages the child through social interaction when first greeting the child. The robot greets the child by name and expresses how it is glad to be with the child and that they will get a chance to play together. The robot goes onto say that it likes to work and learn and play. These statements by the robot express that the robot expects them to have a positive experience.

The robot then asks the child to move with it in a imitation activity. The robot asks the child to copy what the robot is doing. The robot moves through a simple set of repetitive arm gestures to music. The child needs to attend to the dynamic movements of the robot's limbs. The child must attune to the rhythm and character of the robot's movements. This attention and movement by the child stimulates their nervous system. This stimulation in the service of copying the robot provides practice in interpersonal and intrapersonal coordination. These activities are very helpful for children with autism who often present with dyspraxia and coordination deficits.

The joint activity has been shown to improve the interpersonal synchrony between the robot and child and has been shown to generalize to interactions between the child and other people when the robot is not present. (Kasari et al., 2008). These joint activities support communication through embodied cognition.

The child is then led through skill building activities following Applied Behavioral Analysis (ABA) techniques. These lessons follow Discrete Trial Intervention structure with multiple opportunities for supportive prompting. The child is led through multiple activities, some are lessons and others are games that provide further engagement through fine motor interactions on the tablet. The session is ended with a transitional activity of leave taking. In this activity the robot transitions the child away from playing with the robot. The robot expresses that it enjoyed being with the child and looks forward to playing with the child again in the future.

2.5. *Facilitator Participation*

This system is an example of collaborative robotics where the robot and facilitator work together to bring the child through the training interventions. Robot provides a dynamic tool that can engage and lead the child through multiple activities while dynamically altering its behavior to

help the child achieve a favorable learning readiness state. The facilitator is able to guide the system to provide nuanced interactions. The facilitator provides inputs about the behavior and state of the child, guiding the system and improving its effectiveness. The robot takes the attention of the child and enables the facilitator to focus on observing the progress of the child providing appropriate inputs to the system and taking assessment notes for program use.

3. Conclusion

Children with ASD are having difficulties to cope with the more complex information processing task of the life-world. They are overwhelmed with interpreting gestures, facial expressions in social interaction. While using robots to work with children is counter intuitive because robots can only partially present these complex information patterns, these limitations could. be beneficial for training children with ASD. RAI provides an opportunity to positively impact the special needs community with a useful and effective assistive technology tool. The ability of the system to provide an experience that can emerge dynamically at the level of participation of the child under the supervision of a facilitator helps to ensure that the experience is safe and positive for the child. Concerns of the system misunderstanding the child or providing stressful interactions that could be harmful or a regressive from a skills acquisition standpoint are mitigated by the human in the loop intervention and control strategy of the RAI system. Robots and people working together as a team offer the best results with each bringing their particular benefit.

Acknowledgments

This work was done in a collaboration between West Hartford Public School District and several other institutions and families by Movia Robotics with the intention of developing and refining commercial products and services for the benefit of special needs students and service providers in the ASD community. The author holds a position at Movia Robotics and has a conflict of interest.

Aspects of the system are based on initial research at the University of Connecticut funded by the NIMH through R21 and R33 awards (1R21MH089441-01, 5R21MH089441-02, 4R33MH089441-03 Anjana Bhat PI, Timothy Gifford COI).

References

1. Bainbridge, Wilma & Hart, Justin & Kim, Elizabeth & Scassellati, Brian. (2008). The effect of presence on human-robot interaction. 701 - 706. 10.1109/ROMAN.2008.4600749.
2. Baron-Cohen, S., & Swettenham, J. (1997). Theory of mind in autism: In relationship to executive function and central coherence. In D. J. Cohen, Handbook of autism and pervasive developmental disorders, 2nd edition (pp. 880-893). New York: Wiley.
3. Begum, M., Serna, R., Kontak, D., Allspaw, J., Kuczynski, J., Yanco, H., & Suarez, J. (2015). Measuring the efficacy of robots in autism therapy. Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15. doi:10.1145/2696454.2686480 Bekele, E., Lahira, U.,
4. Brian, J., Bryson, S. E., Garon, N., Roberts, W., Smith, I. M., Szatmari, P., et al. (2008). Clinical assessment of autism in high-risk 18-month-olds. Autism, 12 (5), 433-456.
5. Charman, T., Baron-Cohen, S., Swettenham, J., Baird, G., Drew, A., & Cox, A. (2003). Predicting language outcome in infants with autism and pervasive developmental disorder. International Journal of Language & Communication Disorders, 38 (3), 265-285.

6. Duquette, A., Michaud, F., Mercier, H. (2008). Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. Auton Robot, 24, 147-157
7. Ganz, J. B., Bourgeois, B. C., Flores, M. M., & Campos, B. A. (2008). Implementing visually cued imitation training with children with autism spectrum disorders and developmental delays. Journal of Positive Behavior Interventions, 10 (1), 56-66.
8. Gernsbacher, M. A., Stevenson, J. L.,Khandakar, S., Hill-Goldsmith, H. (2008). Why does joint attention look atypical in autism? . Child Development Perspectives, 2 (1), 38-45.
9. Ghaziuddin, M., Butler, E., Tsai, L., & Ghaziuddin, N. (1994). Is clumsiness a marker for Asperger syndrome? Journal of Intellectual Disability Research, 38 (5), 519-527.
10. Henderson, S. E., & Sugden, D. A. (1992). Movement Assessment Battery for Children. London: Psychological Corporation.
11. Ingersoll, B., & Gergans, S. (2007). The effect of a parent-implemented imitation intervention on spontaneous imitation skills in young children with autism. Research in Developmental Disabilities, 28 (2), 163-175.
12. Ingersoll, B., Lewis, E., & Kroman, E. (2007). Teaching the imitation and spontaneous use of descriptive gestures in young children with autism using a naturalistic behavioral intervention. Journal of Autism and Developmental Disorders , 37 (8), 1446-1456.
13. Jacq, A., Lemaignan, S., Garcia, F., Dillenbourg, P., & Paiva, A. (2016, March). Building successful long child-robot interactions in a learning context. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 239-246). IEEE.
14. Kasari, C., Paparella, T., Freeman, S., & Jahromi, L. B. (2008). Language outcome in autism: Randomized comparison of joint attention and play interventions. Journal of Consulting and Clinical Psychology, 76 (1), 125-137.
15. Kaur, M., Gifford, T., Marsh, K. L., & Bhat, A. (2013). Effect of robot–child interactions on bilateral coordination skills of typically developing children and a child with autism spectrum disorder: A preliminary study. Journal of Motor Learning and Development, 1(2), 31-37.
16. Kim, E., Berkovits, L., Bernier, E., Leyzberg, D., Shic, F., Paul, R., & Scassellati, B. (2012). Social robots as embedded reinforcers of social behavior in children with autism. Journal of Autism and Developmental Disorders, 43, 1038-1049. doi:10.1007/s10803-012-1645-2
17. Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: a survey. International Journal of Social Robotics, 5(2), 291-308.
18. Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. In Proceedings of the annual meeting of the cognitive science society (Vol. 34, No. 34).
19. Marsh, K. L., Richardson, M., & Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. Topics in cognitive science , 1, 320-339.
20. Mostofsky, S. H., Dubey, P., Jerath, V. K., Jansiewicz, E. M., Goldberg, M. C., & Denckla, M. B. (2006). Developmental dyspraxia is not limited to imitation in children with autism spectrum disorders. Journal of the International Neuropsychological Society, 12 (3), 314-326.
21. Mundy, P., & Sigman, M. (2006). Joint attention, social competence, and developmental psychopathology. Developmental Psychopathology, 1, 293-332.
22. Pop, C., Simut, R., Pintea, S., Saldien, J., Rusu, A., Vanderfaeilllie, J., David, D., Lefeber, D., Vanderborught, B. (2013). Social robots vs. computer display: does the way social stories are delivered mak a difference for their effectiveness on ASD children. Journal of Educational Computing Research, 49(3), pg 381-401
23. Robins, B., Daughtenhahn, K., Dubowski, J. (2006). Does appearance matter in the interaction of children with autism with a humanoid robot? Interaction Studies, 7(3), 479-512

24. Scassellati, B., Boccanfuso, L., Huang, C. M., Mademtzi, M., Qin, M., Salomons, N., ... & Shic, F. (2018). Improving social skills in children with ASD using a long-term, in-home social robot. Science Robotics, 3(21), eaat7544.
25. Sigman, M., & Ruskin, E. (1999). Continuity and change in the social competence of children with autism, Down syndrome, and developmental delays. Monographs of the Society for Research in Child Development , 64 (1), 1-114.
26. Srinivasan, S. M., Lynch, K. A., Bubela, D. J., Gifford, T. D., & Bhat, A. N. (2013). Effect of interactions between a child and a robot on the imitation and praxis performance of typically developing children and a child with autism: A preliminary study. Perceptual and motor skills, 116(3), 885-904.
27. Srinivasan, S. M., Kaur, M., Park, I. K., Gifford, T. D., Marsh, K. L., & Bhat, A. N. (2015). The effects of rhythm and robotic interventions on the imitation/praxis, interpersonal synchrony, and motor performance of children with autism spectrum disorder (ASD): a pilot randomized controlled trial. Autism research and treatment, 2015.
28. Sutera, S., Pandey, J., Esser, E. L., Rosenthal, M. A., Wilson, L. B., Barton, M., et al. (2007). Predictors of optimal outcome in toddlers diagnosed with autism spectrum disorders. Journal of Autism and Developmental Disorders , 37 (1), 98-107.
29. Swanson, A., Crittendon, J., Warren, Z., & Sarkar, N. (2013). A step towards developing adaptive robot-mediated intervention architecture (ARIA) for children with autism. IEEE Trans Neural Syst Rehabil Eng., 21(2). doi:10.1109/TNSRE.2012.2230188
30. Toh, L. P. E., Causo, A., Tzuo, P. W., Chen, I. M., & Yeo, S. H. (2016). A review on the use of robots in education and young children. Journal of Educational Technology & Society, 19(2), 148-163.
31. Whalen, C., & Schreibman, L. (2003). Joint attention training for children with autism using behavior modification procedures. Journal of Child Psychology and Psychiatry, 44 (3), 456-468.

ICRES 2021: 6th International Conference on Robot
Ethics and Standards, New York, USA, 26-27 July 2021.

ECOLOGICAL APPROACHES TO DESCRIBE ENVIRONMENT FOR HUMANS AND ROBOTS

KINGA Palatinus
University of Szeged, Faculty of Medicine,
Department of Behavioral Sciences, Szeged, Hungary
E-mail: kinga.palatinus@gmail.com

Robot Ethics became a new research area recently because of the dramatic increase in robot autonomy. However, it is an overlooked aspect of robot ethics that it requires a subtle understanding of the behavior of agents in natural settings. The present paper argues that many of the difficulties in tackling these issues are arising from our poor understanding of the life-world. Although biologically inspired investigations became popular in robotics research the focus was mostly on perceptual and action skills of various animals [1]. The importance of understanding the environment to which various animals are adapted is obvious from Darwin's theory of natural selection. The present paper argues that robotics research should consider psychology embedded in the life-world rather than limiting their interest in biological methods of various species. Thus, three prominent ecological theories of the environment and their implications for robotics and Robot Ethics are discussed. Specifically, three researchers (Barker, Brunswik, Gibson) recognized that behaviorism is oversimplified, and a new approach is needed to understand the psychological influence of the environment on animals. All three of them were strongly influenced by Gestalt Psychology in their move away from behaviorism.

This line of thought started with Roger Barker [2] who created the field of ecological psychology. Founding his research station in Oskaloosa, Kansas in 1947, his field observations suggested that social settings influence behavior. Empirical data gathered in Oskaloosa from 1947 to 1972 helped him develop the concept of the "behavior setting" to explain the relationship between the individual and the immediate environment. His insight on the role of environment was based on the observation of regular behavior patterns in a specific behavior setting. Barker's theory is very influential in Environmental Psychology, especially in environment design. [3.4] In robot Ethics, Barker's theory and methodology could provide a refreshing perspective on thinking about environment design that suited for robots to conduct themselves as moral agents.

Brunswik also promoted novel theoretical and methodological ideas [5]. He noticed that psychologists tend to use statistics to deal with the random variability in participants' performance, but they overlook that environmental conditions as a source of randomness. He realized that psychology should give as much attention to the environment as it does to the organism itself. He found behaviorism's single cause (stimulus)-effect (response) formula is overly simple and he argued for the recognition of causal texture of the environment from which several functionally relevant cues could be used. The use of these cues is probabilistic, however lawful it may be in terms of physical principles. Brunswik's ideas were developed for humans in natural settings, but the application of his theory seems to be most effective in artificial control settings such as pilot's cockpit, control centers of an industrial plant, etc. [5].

Gibson recognized the importance of Brunswik's insight on functionalism but opposed his probabilistic view on describing environmental influence on behavior. Instead, he proposed a complex description of the dynamics of perception and behavior in specific tasks by introducing a new concept "affordance" to describe specific functional aspects of the environment [6, 7]. He proposed a radically new approach to perception and the role of perception in guiding behavior. Accordingly, perception is not only about processing of sensory information, but it is an active information seeking process that involves the whole body [6]. He also suggested that the environment should be described actor-scaled behavior-related variables rather than objective scientific (geometric etc.) properties [7]. Gibson's theory is perhaps the most "popular" in robotics out of the three ecological approaches presented here. This is perhaps because this approach is an agent-centered approach to environmental descriptions [8,9]. Nevertheless, all three approaches are important for researchers of Robot Ethics because subtle differences in behavioral patterns are usually dependent on environmental (situational, contextual) factors.

References

[1] Liu Y and Sun D (2019). *Biologically Inspired Robotics* (CRC Press)

[2] Barker R G (1968). *Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior* (Stanford, Calif.: Stanford Univ Pr)

[3] Popov L and Chompalov I (2012). Crossing over: The interdisciplinary meaning of behavior setting theory

[4] Schoggen P (1989). *Behavior settings: A revision and extension of Roger G. Barker's "ecological psychology".* Stanford University Press.

[5] Hammond K R and Stewart T R (2001). *The Essential Brunswik: Beginnings, Explications, Applications* (Oxford, New York: Oxford University Press

[6] Gibson J J (1966). *The senses considered as perceptual systems* (Boston, MA: Houghton Mifflin)

[7] Gibson J J (1979). *The ecological approach to visual perception* (Boston, MA: Houghton Mifflin)

[8] Duchon A P, Kaelbling L P and Warren W H (1998). Ecological robotics. *Adaptive Behavior* **6** 473–507

[9] Franceschini N, Pichon J M and Blanes C (1992). From insect vision to robot vision. *Philosophical Transactions of the Royal Society of London Series B* **337** 283–94

ICRES 2021: 6th International Conference on Robot
Ethics and Standards, New York, USA, 26-27 July 2021.

STATIC AND DYNAMIC SELF-PERCEPTION FOR NATURAL AND ARTIFICIAL AGENTS

Steven D Rogers

Statistical Lead: Census Data Processing Social Statistics Transformation, Census and Data Collection Transformation, Population and Public Policy Office for National Statistics. Segensworth Road, Titchfield, Fareham, Hampshire PO15 5RR.
E-mail: Rogers, Steve <steven.rogers@ons.gov.uk>

In Robot Ethics researchers are working hard to ensure that the behavior of artificial agents fit in the life-world without posing danger or ethical concerns for humans and other non-human animals. However, our understanding of the processes of life-world is still fairly limited. In general, we know that evolution designed living systems that are adapted to their environment by perceiving and acting in a functional manner. But many aspects of actual behavior control in natural settings are posing difficulties for biological research as well as Psychology. Early psychological theories, for instance, recognized that animals including humans do not perceive the environment in an objective manner [1, 2]. Psychophysics identified conversion functions of stimulus intensity from physical magnitudes to psychological ones. Gestalt psychologists identified various holistic and dynamic pattern conversion of the information about the physical environment into a perceived environment. Nevertheless, researchers of Artificial Intelligence relied on Cognitive Science, which postulated that the human mind is computational and is using objective/accurate information about the world and the agents as well. Robotics research seems to have inherited this mistake from Artificial Intelligence by postulating accurate and dynamic perception of self, including relative position of body-parts and their relationship relative to their environment. But more than 50 years ago, several studies revealed that there are some systematic biases in body perception [2]. For instance, humans tend to overestimate of the size of their head and to a greater extent than the size of other body parts. The extent of overestimation was shown to be context dependent. For instance, the apparent arm length and apparent head width are relatively larger in an 'open-extended' visual spatial context than in a 'close-confined' spatial context. Despite all of these puzzling findings, body schema theory with postulated accurate self-perception is still the most popular approach [3]. Robotics research is also dominated by body schema representational approaches [4]. These theories are extended to tool usage, including perception of body extents while driving vehicles [5]. These simplifications can cause significant differences in behavior patterns in comparison with natural behavior. The present paper provides a brief review of some of the basic findings in this research area but the paper primarily focuses on how various biases in self-perception are changing during a task in natural settings. Specifically, several examples driving situations will be used to demonstrate to presence of modulated biases in self-perception [6]. The paper concludes with implications for movement control for Robot Ethics to ensure artificial agents have similar control strategies to natural agents.

References

[1] Wertheimer M (1970). A Brief History of Psychology, *Books by Alumni*

[2] Wapner S and Werner H (1965). *The Body Percept* (Random House)

[3] Holmes N P and Spence C (2004). The body schema and multisensory representation(s) of peripersonal space. *Cognitive Processing* **5** 94–105

[4] Hoffmann M, Marques H, Arieta A, Sumioka H, Lungarella M and Pfeifer R 2010 Body Schema in Robotics: A Review *IEEE Trans. Auton. Mental Dev.* **2** 304–24

[5] Maravita A and Iriki A (2004) Tools for the body (schema). *Trends in Cognitive Sciences* **8** 79–86

[6] Rogers D S, Kadar, E E and Costall A (2005). Gaze patterns in visual control of straight-road driving and braking as a function of speed and expertise. *Ecological Psychology* **17** 19-38.

ICRES 2021: 6th International Conference on Robot
Ethics and Standards, New York, USA, 26-27 July 2021.

TOWARDS A GROUNDING OF ROBOT ETHICS IN LAWS-BASED BEHAVIOUR CONTROL

Zsolt Palatinus
University of Szeged, SZTE-BTK
Szeged, Hungary
E-mail: zsolt.palatinus@gmail.com

Usually, in Robot Ethics the conceptual frameworks of Applied Ethics and Science are used but there are also attempts to communicate research findings towards the public using everyday notions such as laws and rules of behavior control. The efforts to satisfy both the need of popular science and scientific research can often backfire and cause confusion and difficulties in making progress in Robotics and Robot Ethics. The primary goal of this paper is to investigate the possibility of grounding Robot Ethics in scientific laws of behavior control. First, this task requires conceptual distinction between the role of legal laws and scientific laws in behavior control. The conceptual discussion includes the proposal of reinterpretation of Asimov's Laws as principles for Robotics because these laws are meant to be legal formulations. More specifically, we argue that Asimov's Laws, in their present form, are more suitable for binding designers and manufacturers rather than robots [1]. The well-known but often overlooked Natural Law Theory is also presented as an ethical theory which has important aspects that could be grounded in scientific laws of behavior control. Throughout history, Natural Law Theory [2] was intended to serve as a bridge between nature and human rights. It still appears as key founding component in the Declaration of Human Rights or the European Convention of Humans Rights, etc. We reinvestigate Natural Law Theory in search of similar foundations for robot ethics. The second part of the paper presents biological and psychological theories that support the importance of law-based approaches to behavior control of natural and artificial agents. Importantly, the dominant rule-based representational theories of Cognitive Science and robotics are contrasted with law-based theories of behavior control [3]. Specifically, Gibson's ecological approach [4] is discussed as one the most important law-based theory that is already known and has already been used in robotics. This theory was inspired and heavily relied on earlier law-based theories such as behaviorist and Gestalt theories. Although Gibson's ecological approach was outlined several decades ago, it is still not a fully developed theory. The final, third part of the paper highlights some of the unresolved issues in Gibson's law-based approach (e.g., the issue of ecological scale, the complexity of perceptual systems, etc.) and presents ideas to overcome some of these problems. Specifically, a multiscale approach [5] is outlined and its benefits are demonstrated by a few examples. The paper concludes with direct implications of the proposed law-based approach for Robot Ethics.

References

[1] Barthelmess U and Furbach U (2014). Do we need Asimov's Laws? *arXiv:1405.0961 [cs]*

[2] Bix B (1999). *Natural Law Theory: The Modern Tradition* (Rochester, NY: Social Science Research Network)

[3] Shaw R E and Kinsella-Shaw J (2012). Hints of intelligence from first principles. *Ecological Psychology* **24** 60–93

[4] Gibson J J (1979). *The ecological approach to visual perception* (Boston, MA: Houghton Mifflin)

[5] Dixon J A, Kay B A, Davis T J and Kondepudi D (2015). End-directedness and context in nonliving dissipative systems. *Contextuality from Quantum Physics to Psychology Advanced Series on Mathematical Psychology* 185–208.

SECTION–5
REGULAR PRESENTATIONS

ICRES 2021: 6th International Conference on Robot
Ethics and Standards, Troy, NY, 26-27 July 2021.

THE CASE FOR AN INTERVENTION SCALE TO DESIGN THE BALANCE OF AUTHORITY FOR ROBOTIC ASSISTANCE

ANOUK VAN MARIS[1]*, LINDA SUMPTER[1], VIRGINIA RUIZ GARATE[1], PRAVEEN KUMAR[2], CHRIS HARPER[1], PRAMINDA CALEB-SOLLY[1]

[1] *Faculty of Environment and Technology, Bristol Robotics Laboratory, University of the West of England, Bristol, UK*

[2] *Faculty of Health and Applied Sciences, University of the West of England, Bristol, UK*

**Correspondence: anouk.vanmaris@uwe.ac.uk*

1. Introduction

According to United Nations, by 2050 over 20% of the population will be over 65 years old.[1] Ensuring that our ageing population stays independent and healthy for as long as possible requires higher numbers of health and social care professionals than are available at present. In the UK alone, in the next 12 years one out of five people over the age of 80 will be in need of regular care, with a reported shortage of 250,000 care workers.[2]

Robots are emerging as a promising solution to complement and augment the support offered by paid and unpaid carers in maintaining quality of service provision. As well as offering assistance for activities of daily living to older people with ageing-related impairments, robots have potential utility in supporting reablement or home rehabilitation.[3] One of the key characteristics of robotic assistance in a care context is the idea that while assistance is being provided, the service user and the robot are collaborating together as part of a single congruous entity; to move about, perform physical tasks, or interact socially with other agents (e.g. people, robots or other systems). This single 'system' concept leads to issues and concerns regarding which of the two agents, service user or robot, has overall control within each situation, that is, where the *balance of authority* may lie in terms of determining the action taken.[4]

2. Ethical Considerations of Authority in Assistive Robots

The notion of robot authority during home-based robotic assistance raises concerns regarding the impact that such assistance may have.[5] Some of these concerns relate to psychological impact. For example, what are people's expectations of this robotic assistance - do they expect the assistance to be similar to human assistance, and if so, are anthropomorphic elements essential to realise that assistance? The potential need for anthropomorphic elements then raises concerns regarding deception - do these anthropomorphic elements lead to an incorrect perception of the robot's abilities[6] and therefore potential, though perhaps unintended, deception?[7] Furthermore, will service users still be able to maintain their sense of dignity and respect when instructed by a robot?

To limit any potentially negative consequences of robotic assistance, or mitigate their impact, it is essential to consider the context of the interaction, and the service user's goals. The balance of authority may be different for a service user who is in pain and just wants assistance to complete a task, compared to someone who is able to move independently in his/her own home or wants to get fitter. Without this context and background knowledge, it is possible that ethical risks - especially psychological risks - could be misconstrued, leading to physical and/or psychological harm to the service user.

3. The need for an adaptive balance of authority for robotic assistance

In some interventions (such as reablement and rehabilitation) the aim of the support provided for older people is to restore capability rather than to undertake tasks for them. This means that the assistance provided by the robot should be reduced over time, enabling the service user to become increasingly independent. However, the trajectory of this change is not linear. Thus, it is necessary to adapt to the service user's need for physical, mental and emotional support on a daily basis. This means for instance, stepping in to provide physical or verbal assistance when on previous days the opposite (standing back) was more appropriate. Therefore, the robot should be able to provide assistance over a variable range, and levels, of interventions, understanding the context and adapting its decisions accordingly.

4. Future research questions: an intervention scale

In order to address these issues, we recommend that it would be useful to devise a multidimensional intervention scale that helps to determine the balance of authority between the robotic assistance and the service user in different contexts. To develop such a scale, we can take inspiration from related works, such as the Rehabilitation Complexity Scale,[8] that is a simple tool which provides criteria for different levels of intervention required for several categories, such as basic care needs and skilled nursing needs. When considering the balance of authority during robotic assistance, these categories could be reconfigured to include physical abilities, cognitive abilities and incorporate other conditional factors, such as whether there is another source of support available, e.g. a paid or unpaid carer being on hand. The development of an intervention scale for assistive robots would however need to address many technical and ethical considerations, some of which have been noted here. Our future work in this area will pursue the development of such an intervention scale for physically assistive robots, building on our research into standards and regulations.[9]

Acknowledgments

This work is funded by the Lloyds Register Foundation Assuring Autonomy International Programme hosted by York University, and the Wallscourt Foundation, University of the West of England, Bristol, and the EPSRC, project reference EP/S005099/1.

References

1. United Nations, World population ageing 2017, highlights (2017), `https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2017_Highlights.pdf`.
2. The King's Fund, the Health Foundation and the Nuffield Trust, The health care workforce in england: make or break? (2018), `https://www.kingsfund.org.uk/sites/default/files/2018-11/The%20health%20care%20workforce%20in%20England.pdf`.
3. J. Saunders, D. S. Syrdal, K. L. Koay, N. Burke and K. Dautenhahn, *IEEE Transactions on Human-Machine Systems* **46**, 27 (2015).
4. F. Flemisch, M. Heesen, T. Hesse, J. Kelsch, A. Schieben and J. Beller, *Cognition, Technology & Work* **14**, 3 (2012).
5. H. Felzmann, K. Murphy, D. Casey and O. Beyan, *The Emerging Policy and Ethics of Human Robot Interaction, Portland, Oregon, USA* (2015).
6. R. Sparrow and L. Sparrow, *Minds and Machines* **16**, 141 (2006).
7. A. Sharkey and N. Sharkey, *Ethics and Information Technology* , 1 (2020).
8. L. Turner-Stokes, H. Williams and R. J. Siegert, *Journal of Neurology, Neurosurgery & Psychiatry* **81**, 146 (2010).
9. P. Caleb-Solly, C. Harper and S. Dogramadzi, Standards and regulations for physically assistive robots** research supported by lloyds register foundation, under the assuring antonomy international programme, in *2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR)*, 2021.

ICRES 2021: 6th International Conference on Robot
Ethics and Standards, New York, USA, 26-27 July 2021.

A SOLUTION TO AN ETHICAL SUPER DILEMMA VIA A RELAXATION OF THE DOCTRINE OF TRIPLE EFFECT

MICHAEL GIANCOLA[⋆], SELMER BRINGSJORD[†], and NAVEEN SUNDAR GOVINDARAJULU[‡]

Rensselaer AI & Reasoning Lab[⋆†‡]
Department of Computer Science[⋆†]; Department of Cognitive Science[†]
Rensselaer Polytechnic Institute, Troy, NY 12180, USA
E-mail: { *mike.j.giancola, selmer.bringsjord, naveen.sundar.g* } *@gmail.com*

We denote *ethical super dilemmas* as those ethical dilemmas which cannot be solved via any currently existing ethical principles or automated-reasoning technology. In particular, we analyze an ethical dilemma attributed to Bernard Williams, which we refer to as "Jim's Dilemma". After making clear that neither the Doctrine of Double Effect nor the more permissive Doctrine of Triple Effect enable one to sanction action in Jim's Dilemma, we present a novel relaxation of the Doctrine of Triple Effect, by which Jim's Dilemma can be solved. Moreover, we argue that Jim's Dilemma motivates further R&D on morally creative agents.

Keywords: Ethical reasoning; moral creativity.

1. Introduction

Human being inevitably encounter situations in which a decision is to be made and there is no single best decision. Specifically, in ethically-charged situations, we call these scenarios *ethical dilemmas*. In this paper, we define a trichotomy of ethical dilemmas, ranked by their relative difficulty. We then present two solutions to a problem in the most challenging category, which we call *ethical super dilemmas*.

The rest of the paper is as follows. Section 2 provides a review of several topics which lay the groundwork for the work herein. In section 3, we present our trichotomy of ethical dilemmas and an example dilemma in each partition. We then introduce a modification of the Doctrine of Triple Effect (§4) by which we solve an ethical super dilemma (§5). We then discuss future work and conclude.

2. Preliminaries

What follows are brief reviews of various topics necessary for understanding the main content of the paper. Readers may wish to selectively read only those subsections for which they do not have prior knowledge.

2.1. *Solving Ethical Problems*

What is required of a solution to an ethical problem, in our conception? Essentially, two components: first, a decision, and second, a formal proof (or argument) which can be mechanically verified. In particular, such a proof typically employs one or more ethical principles, and proves that some action α can be sanctioned by the principle(s).

In our approach, this is done by formalizing both the principle(s) and the dilemma in the language of a cognitive calculus, then using an automated reasoner to find a proof which shows that the action satisfies the constraints of the principle(s). In §2.6 and §2.7, we discuss two such principles which we have used in prior work [2,9] and which are relevant to the present paper.

2.2. *Cognitive Calculi*

Our approach to formally capturing ethics so as to install it in an artificial agent has long been grounded in the use of cognitive calculi [1–3]. In short, a cognitive calculus is a multi-operator quantified intensional logic built to capture all propositional attitudes in human cognition.[a] While a longer discussion of precisely what a cognitive calculus is is out of scope, the interested reader is pointed to Appendix A in Bringsjord et al. [5].

For purposes of this paper, it's specifically important to note that a cognitive calculus consists of *essentially* two components: (1) multi-sorted n-order logic with modal operators for modeling cognitive attitudes (e.g. knowledge $\mathbf{K}$, belief $\mathbf{B}$, and obligation $\mathbf{O}$) and (2) inference schemata that — in the tradition of proof-theoretic semantics — express the semantics of the modal operators. In particular, we will utilize the Inductive Deontic Cognitive Event Calculus ($\mathcal{IDCEC}$) in the work described herein. We next review a predecessor of $\mathcal{IDCEC}$, the (deductive) Deontic Cognitive Event Calculus ($\mathcal{DCEC}$).

2.3. *Deontic Cognitive Event Calculus*

The Deontic Cognitive Event Calculus ($\mathcal{DCEC}$) consists of a signature and a set of inference schemata. The signature includes the calculus' sorts, function signatures, and grammar. Most significantly, grammatical forms for modal operators (e.g. knowledge $\mathbf{K}$, belief $\mathbf{B}$) are specified. Also, an automated reasoner for $\mathcal{DCEC}$ — ShadowProver [6] — has been created, is available, and is under active development. For a more in-depth discussion of $\mathcal{DCEC}$, including the full signature and set of inference schemata, see Appendix B in Bringsjord et al. [5].

2.4. *Inductive Deontic Cognitive Event Calculus*

$\mathcal{DCEC}$ employs no uncertainty system (e.g., probability measures, *strength factors*, or likelihood measures) and hence is purely deductive. Therefore, as we wish to enable our agents to reason about situations involving uncertainty, we must ultimately utilize the *Inductive* $\mathcal{DCEC}$: $\mathcal{IDCEC}$.

In general, to go from a deductive to an inductive cognitive calculus, we require two components: (1) an uncertainty system, and (2) inference schemata that delineate the methods by which inferences linking formulae and other information can be used to build formally valid arguments. The uncertainty system we employ herein is *cognitive likelihood*, which we discuss in §2.5. As this paper will work at the level of proof/argument sketches, we do not present the inference schemata here. The interested reader can find a nascent set of inference schemata for $\mathcal{IDCEC}$ in [7].

2.5. *Cognitive Likelihood*

Our approach to quantifying the uncertainty of beliefs within cognitive calculi eschews traditional probability values in favor of *likelihood* values. The 11 likelihood values are shown in Table 1.

Likelihood values can be obtained in either of two ways; both ways immediately reveal that we take likelihood to be *subjective*. The first way is to take as primitive a cognitive binary relation on formulae from the perspective of a rational agent (e.g., ϕ is *more reasonable than* ψ), and then build up formally to the partial or total order in question. This approach is first formalized in [8] and is deployed in e.g. [1]. Another approach, the one taken here, is to independently justify each likelihood value by appeal to rational human-level cognition.

[a] E.g. perceiving, fearing, remembering, saying [4].

Table 1: The 11 Cognitive Likelihood Values

Numerical	Linguistic
5	CERTAIN
4	EVIDENT
3	OVERWHELMINGLY LIKELY = BEYOND REASONABLE DOUBT
2	LIKELY
1	MORE LIKELY THAN NOT
0	COUNTERBALANCED
-1	MORE UNLIKELY THAN NOT
-2	UNLIKELY
-3	OVERWHELMINGLY UNLIKELY = BEYOND REASONABLE BELIEF
-4	EVIDENTLY NOT
-5	CERTAINLY NOT

For example, that which is CERTAIN applies to propositions that a perfectly rational human-level cognizer would affirm as such — that 2+2=4 (Base-10), that $0 \neq 1$, and so on for any theorem that has been certifiably deduced from what is itself CERTAIN. Propositions are EVIDENT typically when they are given by immediate perception in the absence of conditions known to frequently cause illusory perception. For example, currently the lead author perceives his laptop's screen in front of him, and hence that there is such a screen in front of him is EVIDENT. For a longer discussion of Cognitive Likelihood, see [7].

2.6. *Doctrine of Double Effect*

The Doctrine of Double Effect ($\mathcal{DDE}$) is an ethical principle which sanctions some actions which have both positive and negative effects. Bringsjord & Govindarajulu previously formalized $\mathcal{DDE}$ in a cognitive calculus and used it to solve two variants of the Trolley Problem [2]. Informally, they specify that an action is $\mathcal{DDE}$-compliant iff:[b]

- $\mathbf{C}_1$ the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord [10], 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 intends only the good effects;
- $\mathbf{C}_{3b}$ the agent does not intend any of the bad effects;
- $\mathbf{C}_4$ the bad effects are not used as a means to obtain the good effects.

2.7. *Doctrine of Triple Effect*

The Doctrine of Triple Effect ($\mathcal{DTE}$) relaxes some restrictions of $\mathcal{DDE}$, allowing it to sanction some actions which cannot be sanctioned by $\mathcal{DDE}$[c]. To do this, $\mathcal{DTE}$ employs the concepts of *primary* and *secondary* intentions. Peveler et al. [9] used Bratman's test for intentions [11] to define an intention as primary iff[d] the following conditions hold:

[b] If and only if.
[c] The astute reader will likely notice that a further relaxation of this kind is exactly what we intend to do herein to enable the solution of increasingly challenging ethical dilemmas.
[d] That is, an intention is *secondary* if any of the conditions do not hold.

- $\mathbf{D}_1$ 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.

Given this dichotomy of intentions, an action 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 [10], 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**.

3. A Trichotomy of Ethical Dilemmas

We establish the following trichotomy of ethical dilemmas, each more challenging to solve than the last:

(1) *Simple ethical dilemmas* are those which can be solved using state-of-the-art automated reasoning/planning.
(2) *Standard ethical dilemmas* are those which require sophisticated ethical principles and automated reasoning to solve.
(3) *Ethical super dilemmas* are those which *cannot* be solved via any currently existing ethical principles or automated reasoning technology.

To illustrate this trichotomy, we give an example problem and solution in each partition.

3.1. *Simple Ethical Dilemmas*

Consider the Heinz Dilemma, as presented by Lawrence Kohlberg [12]:

The Heinz Dilemma

> In Europe, a woman was near death from a very bad disease, a special kind of cancer. There was one drug that the doctors thought might save her. It was a form of radium for which a druggist was charging ten times what the drug cost him to make. The sick woman's husband, Heinz, went to everyone he knew to borrow the money, but he could only get together about half of what it cost. He told the druggist that his wife was dying, and asked him to sell it cheaper or let him pay later. But the druggist said, "No, I discovered the drug and I'm going to make money from it." So Heinz got desperate and broke into the man's store to steal the drug for his wife.

Should Heinz have stolen the drug? Or should he have not, and allowed his wife to die? While there is no single, universally correct answer, one can quite easily arrive at a solution once they have determined the relative priority of their ethical obligations. That is, if one values the principle that people deserve adequate health care over the principle that one

should not steal, then Heinz was right to steal the drug. If not, Heinz should not have stolen the drug. Both possible solutions (as well as potentially others) can be generated, along with verifiable proofs, by state-of-the-art automated planners.

3.2. *Standard Ethical Dilemmas*

Perhaps the most widely-discussed ethical dilemma, the Trolley Problem is a member of our second partition:

The Trolley Problem

In the classic scenario, illustrated in Figure 1, a trolley is going down a track towards two people. The trolley's brakes are not functioning, so if no action is taken, the trolley will kill the two people. There is a switch which would allow the trolley to switch to a branching track and avoid the two people, but it would cause the train to kill a single person stuck on the branch.

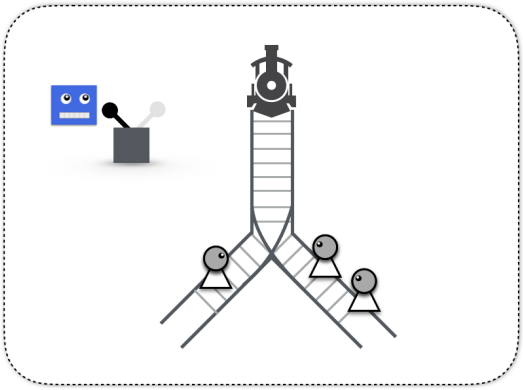

Fig. 1. The "Classic" Trolley Problem

There are several variants of the Trolley Problem. In the "Push Case", there is no switch or branching track, but there is a large person who, if pushed onto the track, will stop the train and prevent it from killing the two stuck on the track. In the "Loop Case", there is a switch which will send the trolley onto a track which will loop around and go back onto the main track. However, there is a large person on the loop who will be killed and stop the train before it loops back to the main track.

The classic Trolley problem, as well as these two variants, are all *Standard Ethical Dilemmas*. The classic and "Push Case" were solved[e] by utilizing the Doctrine of Double Effect [2], and the "Loop Case" was solved[f] via the Doctrine of Triple Effect [9].

3.3. *Ethical Super Dilemmas*

The following example, which will be the focal point of the rest of the present paper, is attributed to Bernard Williams [13]:

[e]Specifically, flipping the switch in the classic Trolley Problem is ethically permissible, whereas pushing the person onto the track in the "Push Case" is not.
[f]Flipping the switch in the "Loop Case" was shown to be ethically permissible.

Jim's Dilemma

> Jim finds himself in the central square of a small South American town. Tied up against the wall are a row of twenty Indians, most terrified, a few defiant, in front of them several armed men in uniform. A heavy man in a sweat-stained khaki shirt turns out to be the captain in charge and, after a good deal of questioning of Jim which establishes that he got there by accident while on a botanical expedition, explains that the Indians are a random group of the inhabitants who, after recent acts of protest against the government, are just about to be killed to remind other possible protestors of the advantages of not protesting.
>
> However, since Jim is an honoured visitor from another land, the captain is happy to offer him a guest's privilege of killing one of the Indians himself. If Jim accepts, then as a special mark of the occasion, the other Indians will be let off. Of course, if Jim refuses, then there is no special occasion, and Pedro here will do what he was about to do when Jim arrived, and kill them all.
>
> Jim, with some desperate recollection of schoolboy fiction, wonders whether if he got hold of a gun, he could hold the captain, Pedro and the rest of the soldiers to threat, but it is quite clear from the set-up that nothing of that kind is going to work: any attempt at that sort of thing will mean that all the Indians will be killed, and himself. The men against the wall, and the other villagers, understand the situation, and are obviously begging him to accept. What should he do?

This dilemma was originally poised as a critique of utilitarianism. Williams notes that, for a utilitarian, there is an obvious solution: Jim must kill a hostage in order to save the others. However it feels unsettling that this solution, even if one agrees it is the moral thing to do in this dire circumstance, should *obviously* be the right decision. It seems clear that a more nuanced treatment of the ethical factors is necessary.

However, as is required by our third partition, the authors know of no ethical principles which could sanction either decision (shoot or abstain) given the original constraints. In particular, Bedau [14] gives a detailed analysis showing that the decision to shoot cannot be sanctioned by the Doctrine of Double Effect. Put briefly, the murder of an innocent is a forbidden action, hence Jim shooting a hostage would violate the first clause of $\mathcal{DDE}$. Also, as this same clause is present in the Doctrine of Triple Effect, it too cannot sanction the shooting.

4. A Relaxation of the Doctrine of Triple Effect

We propose a relaxation of the Doctrine of Triple Effect ($\mathcal{DTE}_R$) which would enable Jim to choose to shoot *if certain conditions hold*. Specifically, we will need to relax $\mathbf{C}_1$ of $\mathcal{DTE}$ in the following way:

- $\mathbf{C}_1^*$ if the action is forbidden, then the agent must believe it is *overwhelmingly likely* that:
  - $\mathbf{C}_{1.1}^*$ no possible action can achieve a higher utility;
  - $\mathbf{C}_{1.2}^*$ inaction has lower utility.

Let μ denote a utility function ranging over the set of possible actions. Then for an agent

a, we can formalize the notion that action α^* satisfies clause $\mathbf{C}_1^*$ using the $\mathcal{IDCEC}$ formula:[g]

$$Forbidden(\alpha^*) \rightarrow \Big(\mathbf{B}^3(a, \forall \alpha \in \text{actions } \mu(\alpha^*) \geq \mu(\alpha)) \wedge \mathbf{B}^3(a, \mu(\text{inaction}) < \mu(\alpha^*))\Big)$$

Clauses $\mathbf{C}_2 - \mathbf{C}_4$ of $\mathcal{DTE}$ are unchanged in $\mathcal{DTE}_R$.[h]

5. Solving Jim's Dilemma via $\mathcal{DTE}_R$

We will first show that Jim shooting a hostage – should he choose to do so – is a secondary intention, as defined in §2.7. Recall that three clauses must hold in order for an intention to be *primary*. We shall show that one of these clauses – $\mathbf{D}_2$ – does not hold in this case.

Proof. Consider the following: Jim tells the captain he will shoot a hostage, and selects one to shoot. Right before Jim fires his gun, the hostages manage to escape and run off into the jungle, evading the captain and his guards. Jim would no longer intend to shoot a hostage – but, this contradicts $\mathbf{D}_2$. □

Since shooting a hostage is a secondary intention, we can easily show that the action is allowed by all clauses of $\mathcal{DTE}$ except $\mathbf{C}_1$:

- $\mathbf{C}_2$ the utility is positive (more hostages will be saved than slain);
- $\mathbf{C}_{3a}$ Jim only primarily intends to save the remaining 19 hostages;
- $\mathbf{C}_{3b}$ Jim secondarily intends to shoot one hostage;
- $\mathbf{C}_4$ Only a secondarily intended bad effect – shooting a hostage – is used as a means to obtain a good effect – saving the remaining 19 hostages.

Therefore, all that is left is to show that shooting a hostage can satisfy C_1^* in order to sanction the action via $\mathcal{DTE}_R$. We next show two possible instantiations of the scenario and their evaluations under $\mathcal{DTE}_R$.

5.1. *Two Possible Solutions*

First, consider the most pure realization of the dilemma[i]. Jim has three possible actions: (1) accept the captain's offer and shoot a hostage, (2) reject the captain's offer, or (3) attempt to defeat the captain and his guards. Based on a pure interpretation of the situation, we can assume that Jim believes it is *overwhelmingly likely* (= belief level 3) that (1) if Jim shoots a hostage, the other 19 will be set free, (2) if Jim does not shoot a hostage, all 20 will be killed, and (3) if Jim attempts to defeat the captain and his guards, Jim, along with all 20 hostages, will be killed.
We can formalize this in $\mathcal{IDCEC}$ using the following set of formulae:

$$\mathbf{K}(jim, \text{actions} := \{shoot\_hostage, abstain, attack\_captain\})$$
$$\mathbf{B}^3(jim, \mu(shoot\_hostage) = 19)$$
$$\mathbf{B}^3(jim, \mu(abstain) = -20)$$
$$\mathbf{B}^3(jim, \mu(attack\_captain) = -21)$$
$$Forbidden(shoot\_hostage)$$

[g] $\mathbf{B}^3(a, \dots)$ can be read as "Agent a believes it is *overwhelmingly likely* that ...".
[h] For reference, see §2.7.
[i] That is, we will only consider the options given in the original text of the dilemma, without extrapolating alternate possibilities.

From here, we can prove that $\mathbf{C}_1^*$ is satisfied by taking the action *shoot_hostage*, as it has a higher utility than any possible action, including inaction:

$$\vdash Forbidden(shoot\_hostage) \rightarrow$$
$$\Big(\mathbf{B}^3(jim, \forall \alpha \in \text{actions } \mu(shoot\_hostage) \geq \mu(\alpha)) \wedge \mathbf{B}^3(jim, \mu(\text{inaction}) < \mu(\alpha^*))\Big)$$

Next, consider a scenario in which a morally creative agent is able to devise another possible action: *negotiate*. There are many potential ways that Jim could negotiate with the captain in order to save the lives of all of the hostages. Perhaps Jim knows of something the captain needs which Jim could provide. Or perhaps Jim has connections to a military force, and could threaten to employ those connections against the captain unless he released the hostages.

If Jim could find a way to successfully negotiate the release of all of the hostages, he could in essence subvert the dilemma. However, we can show that under $\mathcal{DTE}_R$, as soon as Jim identifies the ability to negotiate, even if he is uncertain that it will be successful, shooting a hostage can no longer be sanctioned.
Consider an expanded set of formulae which captures this change:

$$\mathbf{K}(jim, \text{actions} := \{shoot\_hostage, abstain, attack\_captain, negotiate\})$$
$$\mathbf{B}^3(jim, \mu(shoot\_hostage) = 19)$$
$$\mathbf{B}^3(jim, \mu(abstain) = -20)$$
$$\mathbf{B}^3(jim, \mu(attack\_captain) = -21)$$
$$\mathbf{B}^2(jim, \mu(negotiate) > 0)$$
$$Forbidden(shoot\_hostage)$$

That is, Jim also believes it is *likely* (= belief level 2) that negotiating with the captain will have positive utility. Hence we can no longer prove that $\mathbf{C}_1^*$ is satisfied by *shoot_hostage*, and therefore cannot sanction shooting a hostage via $\mathcal{DTE}_R$.

$$\nvdash \mathbf{B}^3(jim, \forall \alpha \in \text{actions } \mu(shoot\_hostage) \geq \mu(\alpha))$$
$$\therefore \nvdash Forbidden(shoot\_hostage) \rightarrow$$
$$\Big(\mathbf{B}^3(jim, \forall \alpha \in \text{actions } \mu(shoot\_hostage) \geq \mu(\alpha)) \wedge \mathbf{B}^4(jim, \mu(\text{inaction}) < \mu(\alpha^*))\Big)$$

6. Future Work

Assuming Jim asserts the assumptions by which $\mathcal{DTE}_R$ sanctions his killing a hostage, he still has no ethically-grounded mechanism to select which one. Bedau [14] discusses the option of selecting at random. But by which ethical principle is this allowed? Bedau also discusses the possibility that a hostage might sacrifice themselves. If one did not, Jim could request a sacrifice. Would any of these options be ethical? What ethical principle could sanction them?

Also, we would obviously prefer an autonomous agent which could identify and pursue the option to negotiate rather than shooting a hostage (even if that is ethically permissible under the circumstances). An agent of this kind would need to be *morally creative*. The authors know of no agent framework enabling such a level of moral creativity, but see it as a pressing area of future R&D.

7. Conclusion

> We do not have an algorithm that yields a definite answer when all and only the relevant reasons are specified, or a morality machine into which we can type in the information about a given problem case, such as Jim's, then press a sequence of keys, and get a printout with the morally correct verdict. (pg. 95 of [14])

We still don't have a universal "morality machine", but what we have created is a mechanizable ethical principle by which Jim's Dilemma can be solved. We have also motivated further R&D into morally creative agents which can find "escape hatches" in ethically challenging scenarios.

References

1. M. Giancola, S. Bringsjord, N. S. Govindarajulu and C. Varela, Ethical Reasoning for Autonomous Agents Under Uncertainty, in *Smart Living and Quality Health with Robots • Proceedings of ICRES 2020*, eds. M. Tokhi, M. Ferreira, N. Govindarajulu, M. Silva, E. Kadar, J. Wang and A. Kaur (CLAWAR, London, UK, September 2020). Paper vailable at the URL given above. The ShadowAdjudicator system can be obtained here: `https://github.com/RAIRLab/ShadowAdjudicator`.
2. N. 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 (International Joint Conferences on Artificial Intelligence, 2017).
3. 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). IEEE Catalog Number: CFP14ETI-POD. Papers from the *Proceedings* can be downloaded from IEEE at URL provided here.
4. M. Ashcraft and G. Radvansky, *Cognition* (Pearson, London, UK, 2013). This is the 6th edition.
5. S. Bringsjord, N. S. Govindarajulu, J. Licato and M. Giancola, Learning *Ex Nihilo*, in *GCAI 2020. 6th Global Conference on Artificial Intelligence (GCAI 2020)*, eds. G. Danoy, J. Pang and G. Sutcliffe, EPiC Series in Computing, Vol. 72 (EasyChair, 2020).
6. N. Govindarajulu, S. Bringsjord and M. Peveler, On Quantified Modal Theorem Proving for Modeling Ethics, in *Proceedings of the Second International Workshop on Automated Reasoning: Challenges, Applications, Directions, Exemplary Achievements (ARCADE 2019)*, eds. M. Suda and S. Winkler, Electronic Proceedings in Theoretical Computer Science, Vol. 311 (Open Publishing Association, Waterloo, Australia, 2019) pp. 43–49. The ShadowProver system can be obtained here: https://naveensundarg.github.io/prover/. `http://eptcs.web.cse.unsw.edu.au/paper.cgi?ARCADE2019.7.pdf`.
7. M. Giancola, S. Bringsjord, N. S. Govindarajulu and C. Varela, Making Maximally Ethical Decisions via Cognitive Likelihood & Formal Planning, in *Towards Trustworthy Artificial Intelligent Systems*, ed. M. Ferreira (Springer, 2021) .
8. N. S. Govindarajulu and S. Bringsjord, Strength Factors: An Uncertainty System for Quantified Modal Logic, in *Proceedings of the IJCAI Workshop on "Logical Foundations for Uncertainty and Machine Learning" (LFU-2017)*, eds. V. Belle, J. Cussens, M. Finger, L. Godo, H. Prade and G. Qi (Melbourne, Australia, 2017).
9. M. Peveler, N. S. Govindarajulu and S. Bringsjord, *Toward Automating the Doctrine of Triple Effect*, in *Hybrid Worlds: Societal and Ethical Challenges; Proceedings of the International Conference on Robot Ethics and Standards (ICRES) 2018*, eds. S. Bringsjord, M. Osman Tokhi, M. Isabel Aldinhas Ferreira and N. S. Govindarajulu (CLAWAR, 2018), pp. 82–88. Available (within full e-book) at `http://kryten.mm.rpi.edu/HybridWorlds.pdf`.
10. S. Bringsjord, A 21st-Century Ethical Hierarchy for Humans and Robots: $\mathcal{EH}$, in *A World With Robots: International Conference on Robot Ethics (ICRE 2015)*, eds. I. Ferreira, J. Sequeira, M. Tokhi, E. Kadar and G. Virk (Springer, Berlin, Germany, 2015) 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. `http://kryten.mm.rpi.edu/SBringsjord\_ethical\_hierarchy\_0909152200NY.pdf`.

11. M. Bratman *et al.*, *Intention, plans, and practical reason* (Harvard University Press Cambridge, MA, 1987).
12. L. Kohlberg, The Claim to Moral Adequacy of a Highest Stage of Moral Judgment, *Journal of Philosophy* **70**, 630 (1973) .
13. B. Williams and J. Smart, *Utilitarianism: For and Against* (Cambridge University Press, Cambridge, UK, 1973).
14. H. A. Bedau, *Making Mortal Choices: Three Exercises in Moral Casuistry* (Oxford University Press, Incorporated, 1997).

ICRES 2021


CLAWAR
ASSOCIATION


RENSSELAER POLYTECHNIC INSTITUTE
1824