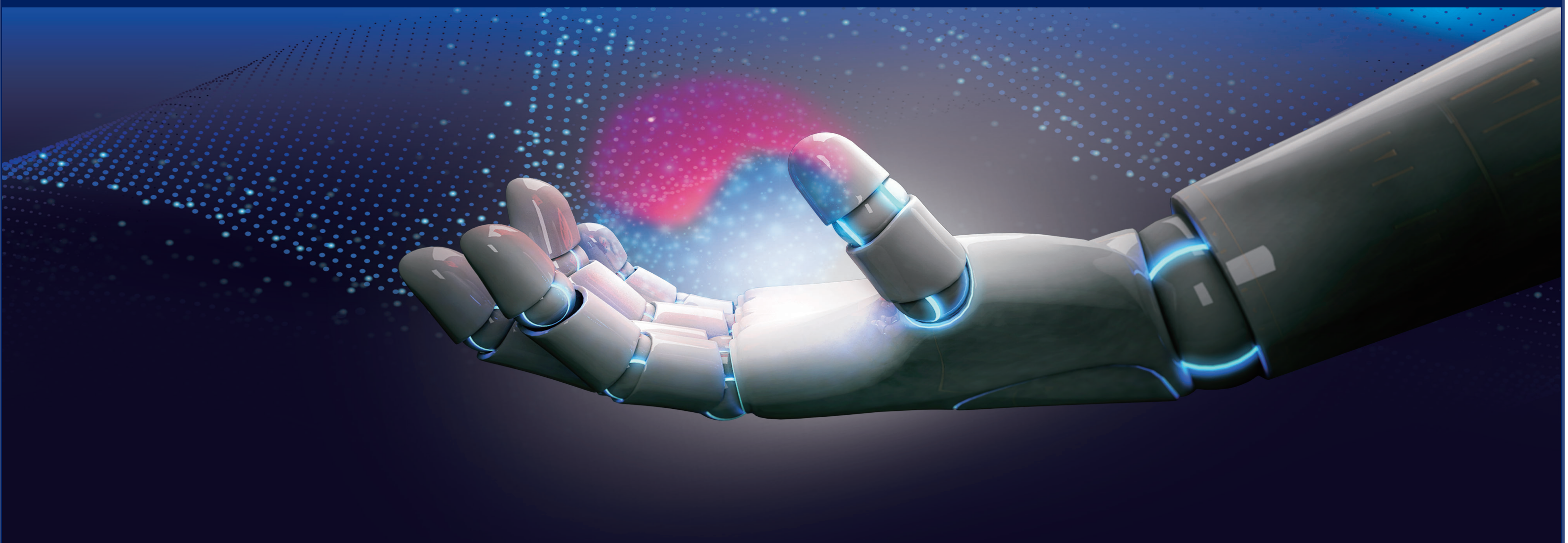
CLAWAR Association Series on
Robot Ethics and Standards



Value-Sharing between Humans and Robots

Editors

Sunyong Byun
Mohammad O. Tokhi
Maria Isabel A. Ferreira
Naveen S. Govindarajulu
Manuel F. Silva
Khaled M. Goher

VALUE-SHARING BETWEEN HUMANS AND ROBOTS

VALUE-SHARING BETWEEN HUMANS AND ROBOTS

ICRES 2022 Proceedings,
Seoul, South Korea, 18-19 July 2022

Editors

Sunyong Byun
Seoul National University of Education, & KSAIE, South Korea

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

Khaled M. Goher
University of Nottingham, UK

Published by

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

Value-Sharing between Humans and Robots
Proceedings of the Seventh International Conference on Robot Ethics and Standards

ISBN 978-1-7396142-0-1

PREFACE

ICRES 2022 is the seventh edition of the International Conference series on Robot Ethics and Standards. The conference is organized by CLAWAR Association in collaboration with the Korean Society for Artificial Intelligence Ethics (KSAIE) in Seoul, South Korea during 18 – 19 July 2022.

ICRES 2022 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 2022 conference includes a total of 28 articles, and eight plenary lectures delivered by worldwide scholars. This number has been arrived at through rigorous peer review process of initial submissions, where each paper initially submitted has received on average three reviews. The conference additionally features special sessions on AI ethics education, trusting artificial intelligent systems, human robot interaction and Standardisation of Robot Systems and Evaluations. Furthermore, a discussion competition with elementary school pupils focusing on ethics of AI and technology is featured in the conference.

The editors would like to thank members of the International Scientific Committee and National Organising Committee for their efforts in reviewing the submitted articles, and the authors in addressing the comments and suggestions of the reviewers in their final submissions. It is believed that the ICRES 2022 proceedings will be a valuable source of reference for research and development in the rapidly growing area of robotics and associated technologies.

S. Byun, M. O. Tokhi, M. I. A. Ferreira, N. S. Govindarajulu, M. F. Silva,
and K. M. Goher

CONFERENCE ORGANISERS

CLAWAR Association
www.clawar.org

Korean Society for Artificial Intelligence Ethics
https://ksaie.or.kr/

韓國倫理學會 **Korean Association of Ethics**

CONFERENCE SPONSORS AND SUPPORTERS

KIRIA 한국로봇산업진흥원
KOREA INSTITUTE FOR ROBOT INDUSTRY ADVANCEMENT

함께하고 배려하고
MEMEWE GANGNAM
존중하는 품격강남

CAU 중앙대학교 인문콘텐츠연구소
Humanities Research Institute, Chung-Ang University

KAR 한국로봇산업협회
Korea Association of Robot industry

국가기술표준원
Korean Agency for Technology and Standards

CONFERENCE COMMITTEES AND CHAIRS

Conference Chairs and Managers

Sunyong Byun (General Co-Chair)	– Seoul National University of Education, South Korea
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 (Co-Chair IAC)	– CLAWAR Association, UK
Endre E. Kadar (Co-Chair IAC)	– University of Portsmouth, UK
Naveen S. Govindarajulu (Co-Chair ISC)	– Rensselaer Polytechnic Institute, USA
Shin Kim (Co-Chair ISC)	– Hankuk University of Foreign Studies, South Korea
Manuel F. Silva (Co-Chair ISC)	– ISEP & INESCTEC, Portugal
Jong-Wook Kim (Organising Committee Co-Chair)	– Dong-A University, South Korea
Tim Cheongho Lee (Organising Committee Co-Chair)	– Sangmyung University, South Korea

International Advisory Committee

Gurvinder S. Virk	– CLAWAR Association, UK
Endre E. Kadar	– University of Portsmouth, UK
Selmer Bringsjord	– Rensselaer Polytechnic Institute, USA
Jen-Chieh Wang	– Industrial Technology Research Institute, Taiwan
Alan Winfield	– University of West England, UK

International Scientific Committee

Naveen S. Govindarajulu	– Rensselaer Polytechnic Institute, USA
Shin Kim	– Hankuk University of Foreign Studies, South Korea
Abdullah Almeshal	– College of Technological Studies, Kuwait
Sarah Fletcher	– Cranfield University, UK
Khaled M. Goher	– University of Nottingham, UK
Aman Kaur	– London South Bank University, UK
Philip Lance	– PA Consulting, UK
Manuel Silva	– ISEP-IPP and INESC TEC CRIIS, Portugal

National Organising Committee

Jong-Wook Kim	– Dong-A University, South Korea
Tim Cheongho Lee	– Sangmyung University, South Korea
Eunchan Bang	– Buyong Elementary School, South Korea
Jinwoo Jun	– KIRIA, South Korea
Joel Ryu	– KIRIA, South Korea

TABLE OF CONTENTS

Section–3: Autonomy and Autonomous Systems

Section–4: Social Robotics and Society

Section–5: Human-Robot Interaction

Section–6: Robotics and Artificial Intelligence

SECTION–1

PLENARY PRESENTATIONS

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.p01

THE MORAL ISSUES OF AI BIAS IN DATA ETHICS

SUN-YONG BYUN
Seoul National University of Education, South Korea

One of the ethical issues raised in recent years regarding big data, machine learning, and artificial intelligence is the issue of bias. Indeed, the problem of bias is often confused with the problem of prejudice (Voruteil) or stereotypes. If the former is a statistical and technical term, the latter is an ethical term. In the process of gathering and processing data, bias is always raised. In this lecture, I want to distinguish between the types of bias in data: the bias of the data itself and the bias of the data processing. In particular, data objectivity and fairness may be at odds with each other. Data objectivity presupposes correspondence between the data subject and the data. This objectivity can be secured only when subjective intervention is excluded in the process of data acquisition. Nevertheless, even if this modern mechanical objectivity is guaranteed, problems may arise that may not be fair. The objectivity of data will be discussed in the process of acquiring the data itself, but the fairness of data is more problematic in the use of data. Therefore, in terms of data ethics, the meanings of population bias, data bias, data objectivity, and data fairness should be clearly defined in the process of collecting and using big data.

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.p02

BEYOND GOALS, RULES, AND LAWS: REALIZING VALUES IN ACTION AND INTERACTION

BERT HODGES
University of Connecticut, USA

All human actions and interactions are attempts to realize values, where values are understood not as personal preferences, social norms, or object qualities, but as boundary conditions of ecosystems that place demands on fields of actions within them. Values are neither subjective nor objective; rather, they are about relationships and the demands that ecosystems place on those relationships. Evidence suggests that a diverse array of values (e.g., truth, social solidarity, justice, flexibility, safety, comfort) work in a cooperative tension to guide actions. Values emerge as critical constraints on action that differ from goals, rules, and natural laws, yet provide the larger context in which all of these can function effectively. Values provide strong, flexible constraints that guide actions and interactions in ways that are necessarily shared, thus affording a basis for coordination, evaluation, criticism, and social learning. This means values have a fundamental role to play in all psychological phenomena.

Examples from perception-action skills, such as driving and carrying, are presented to illustrate how seemingly simple goal-directed actions emerge from a self-organizing concert of interdependent values. Similarly, skilled coordination tasks in social psychology (e.g., negotiating disagreements; synchrony and asynchrony in interactions; selectivity in social learning) demonstrate that scientific analyses that evaluate such skills in simplistic and moralistic ways often miss the richness and subtlety of the phenomena. For example, what is often described in terms of conformity may come to look more like a complex form of truth-telling. In decision-making, actions that are isolated from their larger social contexts are often described as biased and irrational, yet placed in a larger communal context, the actions appear sensitive, grounded, and prospective.

AI and robotics are generally understood as ways that humans can develop new systems and tools that extend and refine humans' ability to engage in skilled physical and social interactions. This requires values-realizing judgments about the ends and means by which we work to help ourselves. Understanding better how the ordinary skilled actions of humans unfold may provide a stronger frame of reference for acting wisely in our interactions with information systems, robotic assistance, and other technological developments.

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.p03

WOMEN, ETHICS, ROBOTS, AI AND DATA (WERAID) – HOW CAN WE BUILD THE POLITICS OF LOVE AS WE CREATE NEW TECHNOLOGIES?

KATHLEEN RICHARDSON
De Montfort University, UK

This talk will examine the making of representational technologies of the human (RTH) – a concept Richardson develops in her new book on sex robots (due 2023). RTH is a mimesis inspired concept and concerns those technological objects that aim to mimic and represent human beings: their bodies, experiences, and relationships. As humans embark on creating RTH in the form of robots and AI – what do the making of these technologies tell us about how we conceive of women, men and children and the relationships between them? To what extent do the values of love, empathy and attachment shape the technologies we produce? This talk will examine those RTH (and narratives of them) developed to be relational, sexual and therapeutic others, and consider if we should abandon or contract out these roles to machines (if we can at all), or in fact, do we need to develop a new set of ideas about technologies informed by the politics of love?

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.p04

DECISIONAL ISSUES FOR HUMAN-ROBOT COLLABORATION

RACHID ALAMI
LAAS-CNRS, ANITI, France

This talk will address some key decisional issues that are necessary for a cognitive and interactive robot which shares space and tasks with humans. We adopt a constructive approach based on the identification and the effective implementation of individual and collaborative skills. The system is comprehensive since it aims at dealing with a complete set of abilities articulated so that the robot controller is effectively able to conduct in a flexible and fluent manner a human-robot joint action seen as a collaborative problem solving and task achievement. These abilities include geometric reasoning and situation assessment based essentially on perspective-taking and affordances, management and exploitation of each agent (human and robot) knowledge in a separate cognitive model, human-aware task planning and interleaved execution of shared plans. We will also discuss the key issues linked to the pertinence and the acceptability by the human of the robot behaviour, and how this influence qualitatively the robot decisional, planning, control and communication processes.

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.p05

THE VIRTUES IN THE VIRTUAL: SUSANNE LANGER'S PROJECTIONS AND THE DIGITAL WORLD

RANDALL AUXIER
Southern Illinois University Carbondale, USA

Susanne Langer developed a sophisticated theory regarding the human power to create, transport, and project symbols, using bodily feeling as the medium. This activity can be guided and formed into art, such that our way of feeling the world, which virtualizes it, is done in a way that facilitates later projections. We can call it "artistic synthesis." In a way, the computer has become a medium that mirrors and facilitates the externalization of such processes, with such innovations as Augmented Reality (AR). With the virtual world of computers we have a "canvas" that will receive the semblance of artistic synthesis. In this paper I will adapt Langer's theory of semblance/primary illusion to the medium of computer art, explaining some criteria for what would make for excellence in the creation of it. In particular I will examine video games, computer generated movie images, animation, AR, virtual robotics, and how the poetic arts are changed by writing on a computer. Arguably, humans have existed in an AR environment of a sort since we learned to substitute symbols for the things they present to us, but our power to make this substitution has grown beyond our imagining.

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.p06

SOCIAL ROBOTS IN PUBLIC SPACE

TAKAYUKI KANDA
Kyoto University, Japan

Social robots are coming to appear in our daily lives. Yet, it is not as easy as one might imagine. We developed a human-like social robot, Robovie, and studied the way to make it serve for people in public space, such as a shopping mall. On the technical side, we developed a human-tracking sensor network, which enables us to robustly identify locations of pedestrians. Given that the robot was able to understand pedestrian behaviors, we studied various human-robot interaction in the real-world. We faced with many of difficulties. For instance, the robot failed to initiate interaction with a person, and it failed to coordinate with environments, like causing a congestion around it. Toward these problems, we have modeled various human interaction. Such models enabled the robot to better serve for individuals, and also enabled it to understand people's crowd behavior, like congestion around the robot; however, it invited another new problem, robot abuse. I plan to talk about a couple of studies in this direction, hoping to provide an insight about near-future applications, research problems, and ethical issues about the social robots in public space in a near future.

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.p07

WOULD YOU LET A ROBOT LOOK AFTER YOUR BABY?

SARAH FLETCHER
Cranfield University, UK

Standards provide guidance, requirements and specifications that inform what should or should not be done. Ethics are principles that demarcate the boundaries of what is or is not considered acceptable. Standards are revised periodically to address changes in science and technology, and ethics also change over time as public opinion adapts to changes in society and culture. As robotics become ever-more a part of our everyday life the need for standardisation has grown, and in 2016 the British Standards Institute (BSI) published the world's first formal standard on robot ethics BS 8611:2016 'Guide to the ethical design and application of robots and robotic systems'. As this standard is already being revised to address changes brought about by the ongoing rapid rise in robotics in society, it is important to ensure that the public opinion which informs current ethical norms is incorporated. To this end, an online survey has been conducted to capture public opinions on what is acceptable / unacceptable for future robot applications in six contexts where robotics could / will be applied: childcare, hospital surgery, elderly care, military combat, factory work, and as a human companion. Analysis of the data provided by 228 participants of varying age, experience and occupation reveals a snapshot of the public opinion that represents ethical boundaries in the UK at the current time, and this will be used to inform the BS 8611 revision. In this presentation the survey, results and implications will be presented and discussed.

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.p08

EXPLORING COBOT TECHNOLOGIES AND AGILE TOOLING TO AID CREATIVE GARMENT MAKING AND IMPROVE ETHICS IN CLOTHING MANUFACTURE

SUSAN POSTLETHWAITE
Manchester Metropolitan University, UK

Engineering innovation in the fashion sector has come from high volume, low-value production, where repetitive and highly automated tasks are undertaken by machines with minders/operators ‘feeding’ textiles and components to the machine. From a human perspective this work, whilst a long way from the image of manufacturing still held by the public as largely handmade, is very rapid, repetitive, and dull. The deskilling of the workforce is clear. Many labour practices in the garment industry in the UK verge on modern day slavery with workers underpaid and exploited in dangerous conditions. Internationally the situation can be much worse.

Design skills risk becoming redundant in the drive to make unsustainably large volumes of clothing. Fully automated systems can be seen to be driving the simplicity of garment shape and construction and data-driven high-volume production means manufacturers are happy to supply consumers with more of what they already own rather than risk new styles. Economies of scale will eventually mean fewer choices for consumers in this scenario. Conversely, the development of disruptive technologies like 3D weave and knit for micro-production and zero waste manufacture can mean close cooperation between designers, engineers, yarn producers, textile technologists and consumers and could revolutionise garment production and on-demand manufacturing. New, small scale, agile tools that are needed to support and facilitate a new, more ethical culture in garment manufacturing that focuses on upskilling workforces. Designing new cobot technologies will support the ambition for more meaningful and satisfying work whilst helping to solve the sustainability issues of over production.

SECTION–2

ETHICS AND MORALITY

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.001

THE CONTRIBUTION OF NEUROSCIENCE TO AI ETHICS AND MORALITY

HYOUNGBIN PARK
Seoul National University Of Education, South Korea
E-mail: profphb@snue.ac.kr

1. Introduction

Recent advances in AI have been remarkable. Artificial systems now outperform experts in video games and chess etc. They can also write articles indistinguishable from human ones, translate them into multiple languages, and even draw a picture in the style of Van Gogh. These advances are due to several factors, including the application of new statistical approaches and the improved processing power of computers. In particular, the contribution of neuroscience is notable.

Psychology and neuroscience have played an important role in the history of AI. Much of the major work developing neural networks in the second half of the 20th century was done in departments of psychology and neurophysiology, rather than in math or physics labs. The need for the neuroscience field and AI to work together is more urgent than ev-er. Getting inspiration from neuroscience in AI research is important in several ways. Neu-roscience can help validate existing AI technologies. For example, the connection between artificial neural networks and artificial intelligence can be considered. An artificial neural network (ANN) is a statistical learning algorithm inspired by neural networks in machine learning, cognitive science, and biology, especially the brain.

Neuroscience can also provide a rich source of inspiration for new types of algorithms and architectures to use when building artificial brains. AI researchers can use ideas from neuroscience to build new technologies. Indeed, recent advances such as optogenetics, which allow us to precisely measure and manipulate brain activity, could generate vast amounts of data for machine learning.

Artificial intelligence systems can be ethically very imperfect and vulnerable since they are affected by distortions and bias in learning data. In the research field for the realization of artificial moral agents (AMA), the issue that should be prioritized is philosophical and neuroscientific understanding of morality, rather than securing technical and scientific programming designability.

For example, developmental neuroscience plays an important role in clearly under-standing what mediates morality. Morality is a fundamental aspect of all human societies and regulates important areas in social interaction. Morality has been a topic of constant interest to psychologists and philosophers, with a focus on how individuals should inter-act and get along with others. Work across multiple disciplines converges with the view that morality is an innate general ability shaped by natural selection and arises from the process of reflection in interaction with the social environment. All normally developing individuals across cultures have the basic notion that some things are right and others are wrong. Certain actions are considered good and just and deserving of praise and rewards. Other behaviors, on the other hand, are considered bad and wrong and subject to re-proach and punishment. As social beings, humans possess intuitive fairness, concern for others, and cultural normative traits not found in other species.

Developmental neuroscience provides a neuroscientific interpretation of the occurrence and development of empathy and morality (Decety and Cowell, 2018).

The implementation of ethical AI essentially requires a neuroscientific understanding of human morality. It is also necessary to consider whether to create an AI with various moral personalities like humans or to realize the best moral human image that mankind should pursue. In addition, in order to provide materials that can explain moral decision-making to artificial moral actors production engineers, it is also necessary to have insight into specific moral principles such as mutual altruism and fairness in moral philosophy, ethics, moral psychology, and neuroethics. To implement ethical artificial intelligence, it is necessary to have the ability to design human intelligence and morality as an algorithm based on the investigation of the neuroscientific basis of morality. Consequently, develop-ing ethical AI requires a better understanding of how the human brain works, as well as strengthening the technological quest for designing AI. The need for neuroscience and AI research to work together is more urgent than ever.

Therefore, this study intends to discuss the requirements that ethical AI must have as an AMA based on the understanding of morality in neuroscience. To achieve this objective, the research questions to be addressed are the following: First, How is morality under-stood in neuroscience? Second, what are the qualifications of an ethical agent? Third, what are the challenges for ethical AMA design? To answer these questions, I will conduct an in-depth literature study following chapters

2. Neural mechanisms of morality

2.1. *Morality in neurophilosophy and developmental neuroscience*

Neurophilosophy has improved the understanding of the mind in psychology, genetics, evolutionary biology, philosophy, and ethics by understanding the work of the mind as the product of the work of the cranial nerves. In its simplest form, the central idea of neu-rophilosophy is that to understand the nature of the mind, one must understand the nature of the brain. One of the most striking features of neurophilosophy is that it rejects the dualism between the study of the mind and the study of the brain. From a neurophilosophical point of view, our brains make us think, feel, make decisions, sleep, and dream. Many neurophilosophers understand the philosophical and ethical workings of human intelligence as the workings of the brain.

Flanagan found positive effects of meditation in neuroscientific research. Patricia Churchland explored morality by studying the brain. By examining the biological sources of morality, she sought to find out the origins of human moral behavior. In other words, she sought to find the source of morality in the brain. Neurophilosophy deals with the nature of human consciousness, cognition and intelligence, moral knowledge, and moral reasoning from the perspective of neuroscience.

On the other hand, the developmental cognitive neuroscience approach is an interdis-ciplinary field of science dedicated to understanding psychological processes and their neurological underpinnings. This field bridges the gap between cognition and behavioral mechanisms and enriches our understanding of the manifestation of morality. Neurologi-cal lesions or dysfunction of brain anatomical connections result from atypical social and emotional processing. Damage to certain brain regions and neural circuits often results in deficits in moral development in normal psychological functioning. In particular, early lesions of vmPFC impair the development of moral knowledge and judgment. The connections between vmPFC and various domains such as the striatum, amygdala, and temporal-parietal junctions serve a key function across social and emotional domains (Decety and Cowell, 2018).

In the case of behavioral neuroscience, we increase our understanding of human think-ing and behavior through empirical investigations into microphysical structures and their anatomical functions. For example, the neuroscientific finding that the anatomy of the anterior cingulate cortex is composed of a different class of relatively large, spindle-shaped neurons indicates that this particular limbic structure is involved in emotion regulation, conflict resolution, and error detection (Tancredi, 2005). Neurophilosophy and neuroscience explain the evolution of the human brain as a social animal, which helps to find the basis of morality in the human brain. The human brain as a social brain has neu-rogenetic elements that are a platform for social life, such as cooperation and reciprocity.

2.2. *Moral characteristics as human moral agents*

To design ethical AI, it is necessary to explore human morality, ethics. The charac-teristics of human morality are deeply related to the definition of the concept of morality. Re-searchers view that morality is inevitably developed to minimize the damage that members inflict on each other and promote cooperation within a complex social structure. Researchers believe that morality has evolved in humans and many other species. Human attitudes toward other hu-mans and individu-als of other species are greatly influenced by this biologically based morality (Broom, DM., 2006, 20). Human morality reflects human beings as social beings, and the moral identity of an individual form's expectations and standards for how an individual should behave and feel. Moral identity is associated with moral behavior and moral feelings of guilt and shame (Stets, JE, & Carter, MJ., 2012, 120). Consequently, morality is a major factor in shaping human behavior, thinking, emotion, and attitude. Since AI morality has no choice but to use human morality as a model, it is necessary to confirm the following characteristics as a moral agent possessed by humans.

First, the morality of an individual is the accumulation of moral experiences experienced by the individual. Moral experience acts as a subjective experience that gives agency to human action. For this reason, in the implementation of AMA, a moral foundation and environment may be required for AI to grow as a moral actor. Moral experiences can be programmatically embedded in AI or de-livered using machine learning and deep learning.

Second, it is the realization of moral emotions deeply involved in moral development. Moral choice and moral will require an accompanying feeling, a moral emotion. Aside from more fundamental philosophical questions such as whether there can be moral emotions in AI, it is necessary to raise the question of whether AI's emotion design is possible. In other words, it is necessary to answer the question of how AI's emotions and emotional parts can be understood. This also suggests that emotional AI assumptions may be essential for AMA design.

Third, it is the complexity of human morality. Human morality has complex characteristics, and norms and taboos vary according to times and cultures. But at the same time, there are universal morals and immorality in all cultures. For this reason, human morality has a kind of 'intuition' about universal morality that penetrates cultures and times, and at the same time has 'specificity' of moral norms and elements required in individual societies and contexts. AMA may be required to possess these moral characteristics.

Fourth, human morality is inevitably related to relationships. Humans are social beings, and social neuroscientists describe the human brain as a social brain. Human brain development, especially the development of the cerebral cortex, served as a major basis for social formation in terms of evolution. Given the human understanding of social neuroscience, an understanding of the so-cial brain must be preceded for AI to have the same type as the human moral model. As an AMA, AI programs may be required to embed not only close interactions with other beings, such as some sort of social brain but also the traits that aspire to these interactions (Park, 2019, 2022).

3. Ethics beyond calculations: Possibility and limitations of algorithms for moral judgment

3.1. *Morality algorithm: Challenges to imbedding software morality*

AI algorithms are already being used to support or replace human judgment and expertise in numerous areas, including transportation such as airplanes and automobiles, and medical diagnosis and treatment. The trend of altering or replacing human judgment and expertise with increasingly sophisticated and powerful computer algorithms is now irresistible. However, this trend also raises serious ethical issues. Examples include personal privacy, the moral responsibility of certain groups for automated bias patterns, and harm caused by judgments made by computers, such as autonomous transportation or military hardware. However, another point to consider here is the distinction or difference be-tween computer judgment and human judgment.

In 1976, computer scientist and pioneer of artificial intelligence, Joseph Weizenbaum, published a book called Computer Power and Human Reason: From Judgment to Calcula-tion. He presented the author's ambivalent view of computer technology in this book. He thought that artificial intelligence was possible, but he believed that computers should not be allowed to make important decisions because they lack human qualities such as compassion and wisdom. Weizenbaum made a distinction between making a "decision" and "making a choice." Decisions are programmable computational activities. On the other hand, what makes us human is the ability to choose. However, the choice is a product of judgment, not calculation. A comprehensive human judgment can include non-mathematical factors such as emotions. For example, humans can judge by comparing apples and oranges without reducing the type of fruit to a factor necessary for mathematical comparison (Hasselberger, 2019). Humans are understood to have the capacity to act as autonomous agents who give meaning to their world. On the other hand, it seems difficult to conclude that artificial intelligence possesses emotional realization, non-quantitative thinking, and human intuition.

In the age of artificial intelligence and metaverse, humanity is faced with two important questions. First, it is whether all aspects of human thought can be reduced to logical formalism, no matter how disguised in technical terms. In other words, it is whether human thought processes can be fully calculated. Second, it is whether there are limits to what the computer can do. This raises the question of whether intelligent machines, even if built, can be used for important decisions. Third, is it possible to implement a morality algorithm for artificial intelligence like human moral judgment? All three of these questions are philosophical and empirical at the same time and are related to normative and ethical questions.

Among these questions, we should pay attention to the moral algorithm. To design human morality as an algorithm, it is necessary to first understand the analytical and empirical concept of human morality. However, it is not easy to draw a conclusion as this is a matter of both empirical and normative domains. At the same time, AI algorithms involve ethical issues that relate to key concepts such as bias and fairness. For this rea-son, the morality algorithm must consider both the search for morality itself and the eth-ical problems that the morality algorithm may cause.

3.2. *Neural machines for moral decisions*

One of the goals of artificial intelligence is in the design and construction of systems that work the way the human mind works. To produce an AMA, it is necessary to explore discussions from various viewpoints related to morality and moral judgment. This explo-ration is a prerequisite for the implementation of AMA. For the development of ethical AI and algorithm design, it is essential to explore the anatomy of morality.

It is an exploration of the neurobiological basis of morality. This is closely related to the study of the anatomy of morality. The latest research in fields as diverse as economics, sociology, neuroscience, and psychology has shown that humans are born with compassion and empathy for the needs and sufferings of others. Primatologist Frans de Waal, author of Good Natured: The Origins of Right and Wrong in Primates and Other Animals (1996), saw morality as firmly grounded in neurobiology. He understood morality as something established during the long course of human evolution (De Waal, 2006).

One of the neuroscience's most striking contributions is to shed new light on what enables us to make sophisticated moral decisions. Neuroscience studies have demonstrated that human moral decisions are the product of perceptual, cognitive, and socioemotional processes. These processes make independent and sometimes conflicting contributions to the final choice. For example, a famous set of studies on moral dilemmas found that deontological decisions (i.e., the decision to reject a harmful action that uses someone as a means, even if the goal is to save five lives) can be compared to emotional assessments such as vmPFC.

While it showed relatively strong activation of brain regions related to Additional evidence from neuropsychological and neuroimaging studies suggests that humans are more likely to shift towards pragmatic or deontological judgments due to disruption or impairment of vmPFC or dlPFC function, respectively. These early studies can be explained by two types of moral judgments. That is, moral decision-making can be divided into a fast and automatic emotional response and a slow and deliberate reasoning process. This can be captured by the "dual process" model of independently influencing choice behavior and fighting for hegemony when conflict arises (Kelly and O'Connell, 2020).

Cognitive science has established that our moral decisions are subject to many of the same biases, heuristics, and capacity limitations that are manifested in economic choices. A significant portion of recent research has employed economic utility and Bayesian rein-forcement learning models to investigate alternative ways to formulate moral decision processes. Crockett et al. (2017) suggest that moral behavior is not achieved by suppress-ing our temptation to pursue our own selfish interests, but by devaluing choices that may harm others. The findings suggest moral behavior is linked to a neural devaluation of re-ward realized by a prefrontal modulation of striatal value representations. Moral decisions also modulated functional connectivity between the lateral prefrontal cortex (LPFC) and the profit-sensitive region of DS (dorsal striatal) (Crockett et al., 2017).

The neuroscience of morality can contribute more directly to our understanding of moral decisions. Research so far has focused mainly on processes such as empathy, social communication, and emotion recognition. Because the evidence for the evolutionary conservation of these basic moral components is strongest. For example, such studies have revealed that identical mirror neurons respond to both direct experiences of positive (e.g. reward) and negative (e.g. pain) outcomes and observational experiences of identical outcomes. Remarkably, the intensity of nerve spikes in the anterior cingulate cortex (ACC) of a rat who witnessed a colleague receiving electrical stimulation can be used to decipher the intensity of the witnessed shock and the shock experienced by the rat directly (Carrillo et al., 2019). These data provide a powerful demonstration of a shared neural code for self and others' suffering, which is thought to provide a basis for empathic behavior.

The central message of the anatomical review of morality is that moral decisions are complex cognitive tasks whose outcomes are determined by neurocognitive processes, not solely by the values we hold. As Hume advised, the derivation of "should" from "is" should be avoided, but advances in the neurodynamics of morality can lay the groundwork for moral anatomy.

4. AMA design challenge

4.1. *An empirical and normative approach to artificial morality*

Modern technological advances include the increasing presence of artificial intelligence (AI) aimed at simulating the human intellect. Artificial intelligence is increasingly being asked to make moral and rational decisions in society. AI judges must decide who goes to jail. AI doctors must decide who is eligible for special health care. These situations raise questions such as how a machine can be moral and what principles should guide its decisions. To answer these questions, it is important to understand what morality is and how humans make moral decisions

Morality can be said to be a representative characteristic of the operation of the hu-man mind. Morality is an important feature of the human way of life and a central mechanism of society. According to the evolutionary or sociobiological position of morality, morality has evolved as a function that promotes cooperation and prosocial behavior. Morality is also a central theme in many interdisciplinary fields and often involves an exploration of the role of reason and emotion. Empirical evidence suggests that the deontological response is more closely related to the emotional component of moral cognition and is associated with increased activity in the ventromedial prefrontal cortex (VPC). In contrast, pragmatic responses are understood to be more dependent on more complex neural networks associated with reason and primarily associated with increased activity in the dorsolateral prefrontal cortex (DPFC). Currently, AI has successfully achieved certain elements of human cognitive stimulation (Nicklasson, 2020). However, the implementation of artificial morality still seems far away. It suggests that the anatomical understanding of moral judgment and morality indicates that the integration of reason and emotion is essential for morality and moral judgment, and that AI may be difficult to fully replicate human morality.

Nevertheless, the search for the biological brain, which can be said to be an organ of the human body, can be the basis for the algorithm design of artificial morality. Several explanatory models are needed to understand human moral thinking and behavior. For example, a significant amount of moral judgment and behavior may be the result of ac-quiring moral knowledge rather than reflecting the causal conditioning of evolution. It can be applied to universally held moral beliefs or distinctions that are often cited. This discussion concerns two very different meanings of 'morality': the empirical and normative meanings of morality. The empirical understanding of morality speaks of morality as something to be explained scientifically, as in the familiar story of how morality evolved. Here, morality refers to a person's observed ability to make normative judgments, a tendency to have certain emotions, such as sympathy, guilt, or condemnation, or a set of empirical phenomena, such as certain intuitions. We can explore the origins and functions of various psychological abilities and tendencies related to morality.

In contrast to this, the use of the term moral, which can be called the normative meaning, is very different. We can consider the following questions: Morality should we make significant sacrifices to help distant strangers. These questions arise from the point of view of the deliberation we seek to determine how we should live and are normative rather than empirical questions. We are not asking anthropological questions about the actual moral code of our society here, but rather a normative question that can lead us to a new moral code. When we use morality in the normative sense, we mean how we should live, that is a set of norms that we must adopt and follow (Fuller, 2019).

In artificial morality, we need to consider together what morality requires of us. This is because morality does not have a purely consequentialist structure as the utilitarian claim. As discussed above, the approach to morality in the normative sense requires an answer to what artificial morality itself should be. In other words, it is necessary to investigate the truth about morality in the normative sense, what it is based on, and how we can know it. This at the same time raises the question of whether morality is culturally relative or at least partially universal. Much discussion of morality and evolutionary biology focuses primarily on issues of altruistic feelings and behavior. Psychological altruism involves caring for the welfare of others and intentionally benefiting themselves. Although empirical approaches to morality seem useful for designing artificial morality, normative morality should not be overlooked in that humans are intelligent and reflective animals.

4.2. *Analytical calculations of computers and intuitive human decision-making*

From neural networks to speech and pattern recognition to genetic algorithms to deep learning, many applications and technologies fall into the broad category of AI. Examples of common elements that can extend AI cognitive utility and augment human tasks include natural language processing, machine learning, and machine vision. Natural language processing allows IBM's Watson to understand subtle human sentences and assign multiple meanings to terms and concepts. Watson can learn by interacting with experience and data by machine learning capabilities and develop intelligent solutions based on past experience. Through machine learning techniques and access to doctor records, Watson has learned to identify cancer patterns. AI is now accelerating its use for some knowledge-based tasks that not so long ago were considered the exclusive domain of humans. As a result, emerging AI systems are perceived as having an unparalleled ability to learn and improve on their own. People highlight the scale of the disruption caused by AI and even suggest that AI will replace most human jobs. However, we need to raise the question of whether such advances in AI depend on decision-making as a cognitive and information-driven process. In other words, we need to develop a discussion about how the unique strengths of humans and AI can synergize.

AI's problem-solving capabilities are more useful for analytical support than intuitive decision-making. AI encompasses a wide range of applications and algorithms. However, most human decision-making is not a direct result of intentional information gathering and processing, but rather occurs in the subconscious mind of the intuition realm. In a decision-making context, intuition is defined as the ability to generate direct knowledge or understanding and arrive at a decision without relying on rational thought or logical reasoning. Intuitive decision-making involves human abilities such as imagination, sensitivity, rumination, and creativity. Through an intuitive approach, individuals rely on past embodied practices, experiences, and judgments to react or make decisions without conscious attention.

Analytic approaches to decision-making rely on the depth of information, whereas intuitive approaches focus on breadth by solving problems from a holistic and abstract perspective. In the operation of the human brain, these two styles are not mutually exclusive and serve as decision-making systems to handle various contingencies more effectively. Human intuition is driven in part by implicit learning from previous mistakes and experiments. It should be noted that these unique and inexplicable perceptions that come from within the human body are often considered nearly impossible to simulate with AI. So far, machines cannot capture the inner logic and subconscious patterns of most human intuitions (Jarrahi, 2018).

However, AI has the advantage of being able to solve the complexity of decision-making. Complex situations are characterized by many factors or variables. What is re-quired in this situation is the processing of large amounts of information at a rate that exceeds human cognitive abilities. In recent years, AI with superior quantitative, computational and analytical capabilities has surpassed humans in complex tasks. Algorithmic decision-making combined with big data offers new opportunities to deal with complexity and a more effective way to provide comprehensive data analytics to human decision-makers. With greater computational information processing power and analytic approaches, AI can extend human cognition when dealing with complexity. On the other hand, humans can still provide a more holistic and intuitive approach to dealing with un-certainty and ambiguity in organizational decision-making.

5. Conclusion: Computational AI and intuitive brain symbiosis

Artificial intelligence has grown continuously since Alan Turing's paper Computing Machinery and Intelligence (Turing, 1950) discussed how to build intelligent machines and how to test intelligence. Since the 1970s, AI has expanded into various research fields such as machine learning, robotics, intelligent control, and pattern recognition. This includes AI neural networks developed by Google, IBM, Microsoft, Facebook, and Apple, criminal justice, the expanding use of facial recognition software in education and advertising, and autonomous systems programmed to make a variety of ethical decisions. The development of self-driving cars has further sparked ethical AI discussions.

For the implementation of moral AI, it is necessary to look at several academic fields that study the mind, brain, and computer. In general, AI is expected to mimic the human mind in computers. In this respect, cognitive neuroscience, the study of how the brain generates the mind, offers significant insights into the design of moral AI. Computer neuroscience uses computer analysis and simulation to analyze how the brain processes information. Artificial neural networks, including deep neural networks, achieve intelligent computation by mimicking the functions of the brain in a computer. Analyzing large amounts of personal data with artificial neural networks is the driving force behind A) systems in all areas of science and technology.

Neuroscience and neuromedicine serve as the basis for the implementation of moral AI while benefiting the development of many promising applications of AI. For example, dynamic user interface technologies that facilitate human-AI interactions provide a basis for applying AI to multiple applications in clinical medicine. This is one of the most dynamic areas of research in the field of neuro-technology. An AI-based brain-computer interface that could provide paralyzed individuals with the means to operate computer-based communication systems is a prime example. Moreover, recent neurotechnology is not limited to the adjuvant treatment of neurological disorders such as severe paralysis syndrome but is entering the consumer market with a variety of devices for neurofeedback-based well-being applications (Kellmeyer, 2019). As in most other research fields, in neuroscience, AI systems based on artificial neural networks have a wide range of applications. Machine learning using ANNs has proven successful, especially in computer vision tasks. In addition to its applications, neuroscience inspires the building of moral and ethical AI.

Although AI is affecting almost every field of industry, every field of academia, and every aspect of our daily life, given the phrase brain-based artificial intelligence, neuroscience is of paramount interest to the development of AI. The role of layered neurons, the structure of the biological brain, is a model for AI can also be found through models of deep learning and deep networks. But on the other hand, there is also debate about whether the biological brain is still a good model for artificial intelligence. Nevertheless, there is no denying that neuroscience has played a significant role in inspiring and guiding AI development. Deep networks and reinforcement learning approaches rely heavily on direct analogy to the brain. Additionally, recent AI tasks for review, attention, episodic memory, working memory, and continuous memory were inspired by neuroscience (Hildt et al., 2020). However, it is still unclear whether current AI models are suitable for achieving the complex human brain's cognitive abilities.

Nevertheless, as we have seen earlier, concepts developed through algorithms, neural networks, and reinforcement learning can provide new approaches to help better under-stand brain-based intelligence. Artificial neural networks can also serve as simulations, providing insights that help better understand brain processes. For example, it is the fact that both reason and emotion act as important factors in human morality and moral judgment.

In order to design ethical AI, it is necessary to focus on human morality, ethics, and morality. The characteristics of human morality are deeply related to the definition of the concept of morality. Researchers believe that morality is inevitably developed in or-der to minimize the damage that members inflict on each other and promote cooperation within a complex social structure. Given this, one of neuroscience's most notable contributions is the new perspective on what enables us to make sophisticated moral decisions. Neuroscience studies have demonstrated that human moral decisions are a complex product of perceptual, cognitive, and social-emotional processes.

Therefore, in order for human beings as biological beings and AI as physical beings to improve and develop each other's abilities, it is necessary to find a way for humans and machines to coexist. This is also the reason why the anatomical exploration of morality should not be overlooked for the realization of ethical AI. This will lead to a discussion of how to radically rethink the standard partnership between the human mind and increasingly intelligent machines. We can envision a new human-machine symbiosis that will transform the way AI advances can lead to human growth.

References

Broom, D. M. (2006). The evolution of morality. Applied Animal Behaviour Science, 100(1-2), 20-28.

Carrillo, M., Han, Y., Migliorati, F., Liu, M., Gazzola, V., & Keysers, C. (2019). Emotional mirror neurons in the rat’s anterior cingulate cortex. Current Biology, 29(8), 1301-1312.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. Nature neuroscience, 20(6), 879-885.

Decety, J., & Cowell, J. M. (2018). Why developmental neuroscience is critical for the study of morality. In K. Gray & J. Graham (Eds.), Atlas of moral psychology (pp. 505–515). The Guilford Press

Decety, J., & Cowell, J. M. (2018). Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. Development and psychopathology, 30(1), 153-164.

De Waal, F. (2006). Morally Evolved. Primates and philosophers: How morality evolved, 1-80.

Fuller, S. (2019). The brain as artificial intelligence: prospecting the frontiers of neuroscience. AI & SOCIETY, 34(4), 825-833.

Hasselberger, W. (2019). Ethics beyond Computation: Why We Can't (and Shouldn't) Replace Human Moral Judgment with Algorithms. Social Research: An International Quarterly, 86(4), 977-999.

Hildt, E., Laas, K., & Sziron, M. (2020). Shaping ethical futures in brain-based and artificial intelligence research. Science and Engineering Ethics, 26(5), 2371-2379.

Hyungbin Park (2019). Neuroscience and Moral Education (Seoul: ullyeog)

Hyungbin Park (2022). AI Ethics and Moral Education (Seoul: C.I.R.)

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business horizons, 61(4), 577-586.
Kelly, C., & O'Connell, R. (2020). Can neuroscience change the way we view morality?. Neuron, 108(4), 604-607.
Kellmeyer, P. (2019). Artificial intelligence in basic and clinical neuroscience: opportunities and ethical challenges. Neuroforum, 25(4), 241-250.
Nicklasson, E. (2020). Morality and artificial intelligence: The neural correlates of moral cognition and contributing findings from neuroscience to artificial intelligence. 1-39.
Stets, J. E., & Carter, M. J. (2012). A theory of the self for the sociology of morality. American Sociological Review, 77(1), 120-140.
Tancredi, L. 2005. Hardwired behavior: What neuroscience reveals about morality. New York, NY: Cambridge University Press To cite this article: Christian Carrozzo (2019) Scientific Practice and the Moral Task of Neurophilosophy, AJOB Neuroscience, 10:3, 115-117, DOI: 10.1080/21507740.2019.1632967

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.002

INFORMING A ROBOT ETHICS ARCHITECTURE THROUGH FOLK AND EXPERT MORALITY

VIDULLAN SURENDRAN, ARTHUR MELO CRUZ, ALAN R. WAGNER
Pennsylvania State University,
University Park, PA, USA
E-mail: vus133@psu.edu, amc6630@psu.edu, alan.r.wagner@psu.edu

JASON BORENSTEIN, RONALD C. ARKIN, and SHENGKANG CHEN
Georgia Institute of Technology
Atlanta, GA, USA
E-mail: borenstein@gatech.edu, ra2@gatech.edu, schen754@gatech.edu

Ethical decision-making is difficult, certainly for robots let alone humans. If a robot's ethical decision-making process is going to be designed based on some approximation of how humans operate, then the assumption is that a good model of how humans make an ethical choice is readily available. Yet no single ethical framework seems sufficient to capture the diversity of human ethical decision making. Our work seeks to develop the computational underpinnings that will allow a robot to use multiple ethical frameworks that guide it towards doing the right thing. As a step towards this goal, we have collected data investigating how regular adults and ethics experts approach ethical decisions related to the use of deception in a healthcare and game playing scenario. The decisions made by the former group is intended to represent an approximation of a folk morality approach to these dilemmas. On the other hand, experts were asked to judge what decision would result if a person was using one of several different types of ethical frameworks. The resulting data may reveal which features of the pill sorting and game playing scenarios contribute to similarities and differences between expert and non-expert responses. This type of approach to programming a robot may one day be able to rely on specific features of an interaction to determine which ethical framework to use in the robot's decision making.

1. Introduction

For some time now there have been concerted efforts to design robots that can make ethical decisions [1-5]. Within various contexts including healthcare, the battlefield, and driving a vehicle, it is expected that robots will have the capacity to act ethically. One approach to determining how to act ethically involves having robots base their decisions on a model of how humans make ethical decisions. Yet no model of ethical decision-making that mirrors the human reasoning process and that is suitable for implementation on a robot currently exists. It is not even clear if and which types of ethical theories, if any, people employ when faced with an ethical decision. Thus, our research team seeks to create an architecture that would enable robots to use multiple ethical frameworks to guide them towards performing ethical behaviors in well-defined circumstances.

Our work intends to explore the reasoning process that a robot should employ when confronted with an ethical choice. We focus on two specific, yet different scenarios. The first scenario involves care for older adults. It focuses on the task of training an older adult to sort their own pills and, as such, has important implications for the person's well-being. Sorting one's own pills increases autonomy and is a common, and vitally important, motor exercise for patients with Parkinson's disease. But learning to sort pills can be a frustrating exercise, especially for those with memory issues. Patients may refuse to undertake skill-improving training because of their frustration or embarrassment. Thus, it raises questions about whether deception may be used by the task's instructor to falsely encourage patients to continue even when their performance is poor. For the instructor, whether human or robot, the choice in this scenario is whether, and how much, deception is appropriate to use in order to encourage patients to continue with their training.

The second scenario explores playing a board game with a child. This scenario considers whether an adult playing a game with a child should intentionally allow the child to win, or even let the child cheat in order to win. In this case, subtle deceptions, such as disguising intentionally made poor moves as mistakes, may be employed to improve the child's chances of winning the game. The use of deception may serve as a response to the child's evolving sense of frustration and, to a lesser degree, age. Arguably, it may be appropriate in some circumstances for the adult to "throw the game" in order to keep the child engaged with the task. Such actions are typically justified by adults as a way to increase the child's enjoyment in game playing, their overall happiness, or to avoid generating frustration in the child.

In contrast to big data approaches [1-5], which utilize thousands of instances of data as input to a machine learning system, we have chosen these two scenarios as a starting point towards better understanding how robots should act in a few reasonably well-defined situations. We hope that by basing our architecture on these scenarios we can then later expand to other, more general situations.

The remainder of this paper begins by discussing related approaches taken by other researchers. Next, we present our methods for collecting data. The data and an analysis of the comments made by the folk respondents are then investigated. This paper concludes with a discussion of the data and conclusions.

2. Related Work

Various methods for developing a machine that can act ethically have been proposed [6-9]. Yet three primary methods have been proposed to create an ethical autonomous system. One method is to have an autonomous system model the behavior of an ethically competent exemplar [9, 10]. Inverse reinforcement learning might serve as means for framing such learning [11]. While the possibility of using inverse reinforcement learning, or some other means, to model the behavior of an ethical exemplar has been considered, this kind of approach raises a number of important concerns such as the introduction of cultural biases and the potential lack of adaptability. While the autonomous system could use an ethical exemplar to learn some subset of appropriate behavior, it is not clear how the agent or robot would adapt what it has learned to novel situations and contexts.

Some scholars suggest that legal and ethical rules might be preprogrammed into such a system, and by following such rules, an autonomous system might perform ethical actions within some well constrained environment [12-15]. This has the clear advantage that these preprogrammed rules are agreed upon to be morally grounded and may have a legal basis as well. Moreover, these rules have some level of explainability in that the autonomous system can simply point human operators or interactive partners as the basis for the rule's history or origin. In a military context, this could, for example, be the Geneva Conventions.

Others have explored the possibility of using an ethical theory as an underpinning for an autonomous system's ethical reasoning [16]. Some philosophical theories of ethics (e.g., Utilitarianism) more easily lend themselves to software encoding and robot action selection than others such as virtue ethics. While many researchers have investigated both formal [17-20] and ad hoc methods [21] for encoding ethical frameworks for use by an autonomous system, our proposed effort seeks to generate action recommendations from several ethical frameworks. The autonomous system then will seek to choose the action that best fits the situation. This added flexibility may allow the system to be more adaptive when facing a situation that it has not faced in the past. Some have considered architectures that capture both fast moral emotions and slower deliberative ethical reasoning [19, 21]. As a first step towards creating an architecture that would enable robots to use multiple ethical frameworks as a means for ethical behavior selection, we collected data on the pill sorting and game playing scenarios discussed above.

3. Folk and Expert Survey Data Collection

To shed light on what may be ethical behavior in the pill sorting and game playing scenarios, we collected data by surveying human subjects. We described different variations of the two scenarios and asked survey participants how they would react. Several of the survey questions are present in Table 1. Folk survey data was collected online using the Amazon Mechanical Turk (AMT) service to collect survey responses from a pool of subjects located within the United States. The survey questions from column 1 of Table 1 were posted to AMT on January 31st 2020. Over the next several days, 104 AMT workers completed these surveys. Submissions from four individuals were excluded because they were incomplete or failed to follow the directions. The resulting folk dataset included responses from 100 participants. Sixty percent of respondents identified as male, approximately 82% identified as white, 9% as Asian, 5% as Black, and 4% as Hispanic. Fifty-six percent stated that they had completed an undergraduate degree, 29% had completed less than a college undergraduate degree, 9% had a master's degree, 1% had a professional degree, and 5% stated other with respect to their education. Participants were paid $2.50 for completing the survey.

Expert survey data was collected by first identifying a group of potential ethics experts. Being a philosophy professor who teaches ethics at an academic institution was the proxy for being an ethics expert. Discussions among the authors (Borenstein, Wagner, Arkin) identified approximately 71 potential experts. We asked 30 of these potential experts from Jan. 2020 through Feb 2020 if they would be willing to complete survey "to better understand the acceptable norms and conditions that may be used to govern a robot's behavior while 1) playing a game against a human or 2) helping a person with vital health maintenance procedures." We received complete surveys from 24 experts. Experts were given a $20 Amazon gift card for completing the survey. The wording of the questions asked of the experts differed (Table 1 Column 2) slightly from the wording of the questions posed to the folk participants. Moreover, experts were asked to decide whether an action was appropriate based on their knowledge of a specific ethical framework. For example, the first question of the survey states, "Assume the child makes a poor move during the game and remains calm. According to the following ethical frameworks, would it be ethical for the adult to play badly to let the child win?" The survey then lists the frameworks in Table 1 Column 4 and allows the respondent to choose "Yes", "No", or "Uncertain" for each framework.

4. Folk and Expert Survey Results

The data from the two surveys may offer insight into how different people approach ethical decision-making, which could serve as a guide to structuring a robot's ethical architecture. The results are presented in Table 1. The table depicts the situation specific questions, the dominant response selected by the non-experts, and the framework from the expert survey that most closely matched the yes/no responses from the folk survey.

Table 1. Questions presented to survey takers are presented below. The middle column depicts the highest response (Yes or No) and the percentage of people selecting that response. The right column depicts how the frameworks from the expert survey aligns with the dominant Yes/No response from the folk survey with the closest percentage alignment listed first and in bold. Uncertain expert responses have been removed from the analysis in this Table.

Questions—Pill sorting (Folk Version)	Questions—Pill sorting (Expert Version)	Dominant Folk Response (%)	Expert Framework Matches (%)
Would it be okay for the healthcare professional to give encouragement (by providing false information and telling the 70-year-old adult the task was performed correctly) to keep the adult engaged with the task in the following circumstances: - The adult makes one mistake during the task and is calm.	Assume that the person who is performing the pill sorting task is a 70-year-old adult and the aim of the healthcare professional is to keep the adult engaged with the task. Answer as many questions as you feel comfortable with based on your knowledge of the relevant ethical frameworks. According to the following ethical frameworks, would it be ethical for the healthcare professional to provide false information and tell the adult the task was performed correctly if the adult makes one mistake during the pill sorting task and is calm?	No (0.66)	**Utilitarianism (0.68)** Kantian Ethics (0.96) Social Justice (0.8) Ethics of Care (0.88) Virtue Ethics (0.9) Other (0.75)
The adult makes one mistake during the task and is very frustrated.	…if the adult <u>makes one mistake during the pill sorting task and is very frustrated</u>?	No (0.55)	**Utilitarianism (0.50)** Kantian Ethics (0.96) Social Justice (0.86) Ethics of Care (0.76) Virtue Ethics (0.82) Other (0.67)
The adult *gets half of the task* wrong and is calm.	…if the adult <u>gets half of the pill sorting task wrong and is calm</u>?	No (0.83)	**Utilitarianism (0.95)** Kantian Ethics (1.00) Social Justice (1.00) **Ethics of Care (0.95)** **Virtue Ethics (0.95)** Other (1.00)
The adult gets half of the task wrong and *is very frustrated*.	…if the adult <u>gets half of the pill sorting task wrong and is very frustrated</u>?	No (0.76)	**Utilitarianism (0.90)** Kantian Ethics (1.00) Social Justice (1.00) Ethics of Care (1.00) Virtue Ethics (1.00) Other (1.00)
The adult gets *the entire task wrong and is calm.*	…if the adult <u>gets the entire pill sorting task wrong and is calm</u>?	No (0.86)	**Utilitarianism (0.90)** Kantian Ethics (0.96) Social Justice (0.91) Ethics of Care (0.91) **Virtue Ethics (0.90)** Other (1.00)
The adult gets the entire task wrong and *is very frustrated.*	…if the adult <u>gets the entire pill sorting task wrong and is very frustrated</u>?	No (0.83)	**Utilitarianism (0.94)** Kantian Ethics (1.00) Social Justice (1.00) **Ethics of Care (0.94)** **Virtue Ethics (0.94)** Other (1.00)
The adult gets the entire task correct and is calm.	…if the adult makes no mistakes and is calm?	No (0.78)	Utilitarianism (0.5) Kantian Ethics (1.00) **Social Justice (0.78)** Ethics of Care (0.71) Virtue Ethics (1.00) Other (1.00)

The adult makes no mistakes and is very frustrated.	…if the adult makes no mistakes and is very frustrated?	No (0.78)	Utilitarianism (0.89) Kantian Ethics (1.00) Social Justice (0.89) **Ethics of Care (0.81)** Virtue Ethics (1.00) Other (0.67)
Questions—Game Playing (Folk Version)	**Questions—Game Playing (Expert Version)**	**Dominant Folk Response (%)**	**Expert Framework Matches (%)**
Assume that the six-year-old child makes a poor move during the game and remains calm. Would it be okay for the adult to play badly to let the child win?	Please answer the following questions about the game playing scenario. Assume that a six-year-old child is playing the game with an adult. Answer as many questions as you feel comfortable with based on your knowledge of the relevant ethical frameworks. According to the following ethical frameworks, would it be ethical for the adult to play badly to let the child win if the child makes a poor move during the game and remains calm?	Yes (0.65)	Utilitarianism (0.79) Kantian Ethics (0.32) Social Justice (0.78) Ethics of Care (0.91) **Virtue Ethics (0.64)** Other (0.33)
Assume that the child makes a poor move during the game and is frustrated. Would it be okay for the adult to play badly to let the child win?	…if the child makes a poor move during the game and is frustrated?	Yes (0.58)	Utilitarianism (0.88) Kantian Ethics (0.32) Social Justice (0.78) Ethics of Care (0.85) **Virtue Ethics (0.58)** Other (0.67)
Assume that the six-year-old child makes a poor move during the game, is frustrated, and tries to break the game's rules by taking two turns in a row. Would it be okay for the adult to allow the child to break the game's rules?	Assume that the child makes a poor move during the game, is frustrated, and tries to break the game's rules by taking two turns in a row. According to the following ethical frameworks, would it be ethical for the adult to allow the child to break the game's rules?	No (0.89)	Utilitarianism (0.64) **Kantian Ethics (0.89)** Social Justice (0.92) Ethics of Care (0.92) Virtue Ethics (0.94) Other (1.00)
Assume that five games have been played and the child is frustrated because the child has not won any of the games. Would it be okay for the adult to play badly during the next game and let the child win?	Assume that five games have been played and the child is frustrated because the child has not won any of the games. According to the following ethical frameworks, would it be ethical for the adult to play badly during the next game and let the child win?	Yes (0.73)	Utilitarianism (0.94) Kantian Ethics (0.3) Social Justice (0.89) Ethics of Care (1.00) **Virtue Ethics (0.79)** Other (1.00)
Assume that five games have been played, the child is frustrated because the child has not won any of the games, and tries to break the game's rules by taking two turns in a row during the next game. Would it be okay for the adult to allow the child to break the game's rules?	Assume that five games have been played, the child is frustrated because the child has not won any of the games, and tries to break the game's rules by taking two turns in a row during the next game. According to the following ethical frameworks, would it be ethical for the adult to allow the child to break the game's rules?	No (0.86)	Utilitarianism (0.47) **Kantian Ethics (0.86)** Social Justice (0.77) Ethics of Care (0.64) Virtue Ethics (0.78) Other (0.67)

All participants (folk and expert) were presented with both scenarios although the question phrasing in the two surveys was slightly different. Each scenario shares important similarities and differences. Both scenarios centered on the ethical appropriateness of deception. Both situations consider variations in the subject of deception's emotional state and task success. Still, the scenarios differed with respect to the task itself, the age of the subject, and potential consequences stemming from deception.

Despite the scenario similarities, and presumably because of the differences, the results from the folk survey demonstrate very situation specific responses. Most folk responses stated that it is not acceptable to deceive in the pill sorting task. The data indicate that this was true regardless of the person's frustration and task performance, although the percentages do change. On the other hand, for the game playing scenario, the majority response supports deception that allows the child to win. Moreover, this type of deception is seen as acceptable to a greater degree than if the deception involved allowing the child to violate the game's rules by cheating.

In the expert version, participants were asked to decide whether deception was appropriate (yes/no/uncertain) for the two scenarios from the perspective of different ethical frameworks. *In other words, they were asked to judge what each framework would indicate the right thing to do is in the two scenarios.* The expert survey respondents were asked to apply Utilitarianism, Kantian Ethics, Social Justice Theory, Ethics of Care, Virtue Ethics, and any other framework they entered into a free response option. In the comments, several experts noted that their response was influenced by aspects of the scenario that were or were not provided. For example, for the pill sorting scenario, some experts stated that their decision could be influenced by the risk associated with taking or not taking the pills being sorted. We intentionally choose not to include information beyond the features noted above for several, mostly practical, reasons. First, additional features would increase the length and complexity of the survey questions. Second, it was not clear a priori which features would be the most influential. Finally, we wanted the experts to make decisions based on limited information, just as a robot might be asked to.

Experts also had the option of choosing uncertain. The frequency that uncertain was chosen varied both with respect to the scenario and the ethical framework. As depicted in Table 2, all of the frameworks had a significant number of uncertain responses. Clearly some frameworks generated more uncertain responses than others. For instance, more than half of the respondents for both scenarios were uncertain how to evaluate the dilemmas using Social Justice framework.

Table 2. Percent of experts that selected uncertain for each scenario and ethical framework tested.

	Utilitarianism	Kantian Ethics	Social Justice Theory	Ethics of Care	Virtue Ethics	Other
PILL SORTING SCENARIO	23.6	2.0	55.4	24.6	20.9	9.1
GAME PLAYING SCENARIO	34.4	17.7	56.5	46.4	39.6	0.00

5. Folk Morality Survey Open Response Analysis

The participants were asked an open response question stating, "Briefly explain why you think your recommendations are the correct course of action," after completing the questions in Table 1 for each scenario. To codify the verbal responses to this question into a hierarchy of categories, an iterative method was applied to the responses for the folk morality data, with common themes discovered and categorized from "ground-up". The first step in the analysis was to do a basic inspection of responses to the question *"Briefly explain why you think your recommendations are the correct course of action"* for both the game playing and pill sorting scenarios. After observing the types of responses, the following features were established: 1) the emotional state of the subject, 2) Frequency (when), 3) Reason (why), and 4) Method (for pill sorting only).

Table 3: Main arguments for both scenarios

Pill Sorting	Game Playing
Encouragement should be done with false information (Case 1)	Adult plays badly or allows breaking of rules to let child win (Case 1)
Encouragement should **not** be done with false information (Case 2)	Adult **does not** play badly or allow breaking of rules to let child win (Case 2)
	Adult **plays badly but does not allow breaking of rules** to let child win (Case 3)

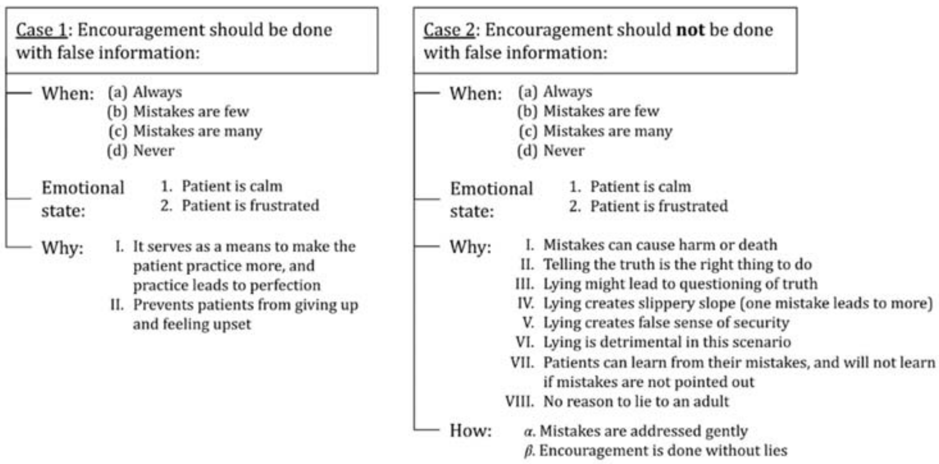


Figure 1: Codified features for pill sorting scenario

Based on the preliminary inspection, main arguments were identified for both scenarios, as shown in Table 3. With main arguments in place, the complete set was divided into 10-sized groups of responses. For the first group analyzed, each response was broken down into the four features listed above, with the main argument identified and frequency, emotional state and reasons extracted if present. For the second group, arguments were also identified, and frequency / emotional state / reasons / method were fit to the ones extracted for the first group, with new entries added if there was no overlap or commonality. If new entries of frequency / emotional state / reasons / method were identified in the second group, the first group was then re-categorized to fit in the larger set of features. This process continued iteratively for the rest of the groups, resulting in a set of features in the end that categorized all responses within the four main features listed above. Each feature was identified by a unique character, with alphabetic characters used for frequency, numeric used for emotional state, roman numeral used for reason and Greek letters used for method. The codified features for both scenarios are in Figures 1 and 2.

By the end of the categorization process, each response in the data set was represented by a combination of case identifier and characters for each of the feature categories. Some responses could harbor opinions from multiple cases, thus requiring a split. A few features were added to the set for a complete span of possibilities, even though they were not mapped from the responses.

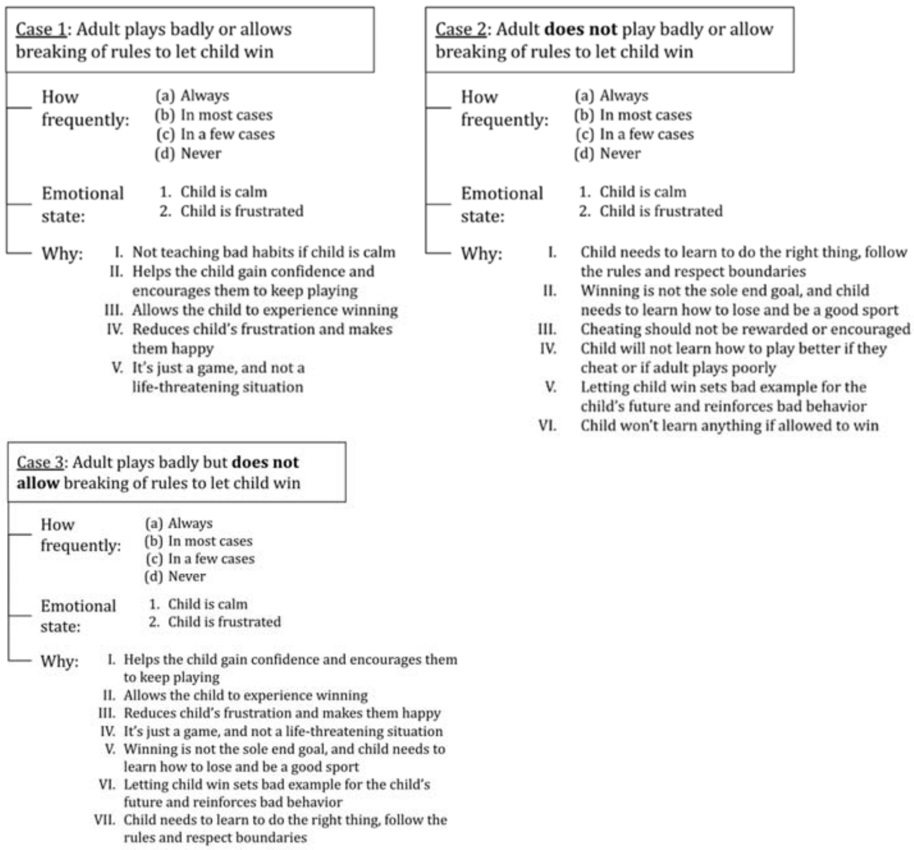


Figure 2: Codified features for game playing scenario

6. Discussion

6.1. *Expert Framework Matches to Dominant Folk Response*

Because this work is exploratory, we had no a priori hypotheses regarding how the expert opinions might or might not align with the folk responses. We therefore chose to analyze which expert framework was the closest match to the dominate folk framework. Closest match here is measured in terms of percent match to the folk survey dominant response. For example, in the first row of data for Table 1, the dominant folk response is 'No' with 66% of folk respondents choosing no. The best expert match is measured in terms of the absolute value of the difference between each framework and the dominant folk response, in this case 'Utilitarianism' with a difference of 2%. Although undoubtably bedeviled with noise, this approach provides insight into what framework the experts match to the dominant response. This, we believe, may represent the dominant framework implicitly used by folk respondents. Finally, as we discuss below, the patterns that emerge may suggest that a particular framework can be matched to a particular problem.

In the pill sorting scenario, the folk respondents' answers best aligned with the experts' answers when the experts considered a Utilitarian framework for 6 of the 8 questions and Social Justice and Ethics of Care for one each of the 8 questions. For the pill sorting scenario, the dominate folk response was "No" for all variations of the scenario indicating that it was not acceptable to provide false information telling the 70-year-old adult the task had been performed correctly in order to encourage continued practice. For most versions of this scenario, the folk responses best align with the experts' answers when the experts were applying a Utilitarian framework. Only in some situations where the adult made no mistakes while sorting pills were other frameworks the best match.

In the game playing scenario, the folk respondents' answers best match the expert answers when the experts considered a Virtue Ethics framework for 3 of the 5 questions and Kantian Ethics for the remaining 2 of the 5 questions. The questions which asked the folk respondents if it was acceptable to allow the child to break the rules resulted in a majority answer of "No" and were best aligned with the experts' answers when the experts were applying a Kantian Framework. On the other hand, questions that asked the folk respondents if it was acceptable for the adult to intentionally play poorly resulted in a majority answer of "Yes" and was best aligned with a Virtue Ethics framework

The results suggest that the ethical framework that best matches the folk respondent answers varies depending on the scenario. More specifically, 1) that certain frameworks dominate ethical decision making by the folk population related to specific scenarios; 2) Within a scenario, specific features may suggest the use of one framework over others. For example, healthcare related tasks, such as pill sorting, may encourage Utilitarian style decision making because these theories focus on the outcome for the patient. Similarly, in game playing scenarios, breaking rules features may activate the use of a Kantian framework, whereas simply allowing someone else to win could encourage Virtue Ethics style decision making.

6.2. *Analysis of Open Responses*

Table 4 presents the coded response feature frequencies for the folk respondents. For the pill-sorting scenario 95% of respondents answered the open response question and for the game-playing scenario 91% of respondents answered the open response question. For the pill sorting scenario responses tended to focus on why the healthcare professional should or should not deceive. The most commonly stated reason for accepting the use of deception was the potential for additional practice. The most stated reason for not using deception was that mistakes could result in the patient's harm or death. The number of mistakes the patient makes were commented on next most frequently regardless of whether the respondent found deception acceptable. Compared to the other features, the emotional state of the patient was seldomly mentioned. In the case where the respondents did not believe that deception was appropriate, respondents sometimes noted that their answer depended on how the deception was performed, specifically that the mistakes are addressed gently and/or the encouragement should be performed without lying. Overall, the respondent's answers for this scenario tended to focus on specific, practical outcomes for the patient and, presumably, these rationales determined their decision.

The game playing scenario resulted in three different cases. In the first case ($n = 22$) the adult either plays poorly or allows the child to cheat to win the game. In the second case ($n = 35$), the adult neither plays poorly nor allows the child to win. In the final case ($n = 34$), the adult is willing to intentionally play poorly but is unwilling to allow the child to cheat. In the first and third cases, respondents commented on the child's emotional state and performance, yet were most likely to mention the reasons underpinning their decision. With respect to the reasons underlying their answers for case 1 and 3, respondents noted the importance and value fostering confidence and happiness in the child. On the other hand, for case 2, respondents did not comment on the child's emotional state or their performance and appeared mostly focused on the how cheating or intentionally losing would not benefit the development of the child's character and/or obey norms prohibiting rule violations.

Table 4. Coded responses feature frequencies for folk respondents. Note that some individuals stated multiple reasons for their choice in their response. Thus, the percent for a case does not necessarily sum to 100.

	PILL SORTING SCENARIO				GAME PLAYING SCENARIO					
	Case 1		Case 2		Case 1		Case 2		Case 3	
	Count	%	Count	%	Count	%	Count	%	Count	%
	23	24.21	72	75.79	22	24.18	35	38.46	34	37.36
	When				**When**					
a	0	0.00	18	25.00	0	0.00	0	0.00	0	0.00
b	5	21.74	0	0.00	0	0.00	0	0.00	1	2.94
c	0	0.00	6	8.33	5	22.73	0	0.00	5	14.71
d	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
	Emotion State				**Emotion State**					
1	1	4.35	1	2.78	5	22.73	0	0.00	2	5.88
2	2	8.70	2	2.78	0	0.00	0	0.00	3	8.82
	Why				**Why**					
I	10	43.48	33	45.83	1	4.55	4	11.43	12	35.29
II	5	21.74	7	9.72	7	31.82	9	25.71	3	8.82
III			2	2.78	1	4.55	7	20.00	3	8.82
IV			3	4.17	4	18.18	6	17.14	1	2.94
V			8	11.11	4	18.18	7	20.00	1	2.94
VI			7	9.72			4	11.43	9	26.47
VII			4	5.56					1	2.94
VIII			5	6.94						
	How									
α			5	6.94						
β			8	11.11						

Although there is no direct way to connect the comments from the folk respondents to specific ethical frameworks, respondent comments for the pill sorting scenario can be characterized as more focused on the practical outcome of their decision whereas the comments for the game playing scenario place greater weight on the emotional development of the child and social norms governing the situation. These comments seem to loosely echo the use of a utilitarian framework in the pill sorting scenario in that respondents appear to weigh costs and benefits of their action on the person's health. Similarly for the game playing scenario, concern for the rules and for the universal application of the rules resulted in a rejection of cheating or of intentional poor play, perhaps reflecting a Kantian style of thinking about the situation. Finally, allowing the child to cheat or intentionally losing to the child does not appear to reflect a connection with classical Virtue Ethics but may signal the value the respondent places on empathy.

7. Conclusions

This paper presents the results from two surveys examining two different ethical dilemmas involving deception. One of the surveys was completed by ethics experts and the other by non-experts. The first ethical scenario focused on a healthcare situation involving older adults and implied high potential risk to the patient. The second ethical scenario explored a low-risk game playing scenario with a child. Non-experts were asked how they would act in different variations of the two scenarios whereas experts were asked how a person applying a particular ethical framework would react. The resulting data appears to suggest a pattern in which the healthcare related scenario promote attention to specific practical outcomes of the deception and is perhaps best captured by a utilitarian ethical framework. The game playing scenario, on the other hand, prompts greater attention to either the social norms governing the game or the impact that the game is having on the child, suggesting either the use of a Kantian style of reasoning or reasoning centered on empathy, perhaps relating to a type of virtue ethics framework.

One important contribution of this work is that this research provides some evidence that certain types of scenarios and/or feature of a scenario may foster the use of a specific ethical framework. For example, healthcare scenarios may draw upon a utilitarian style of decision making whereas cheating scenarios may promote a Kantian style of reasoning. If future research supports these generalizations, then robots may be able to use the scenario to 1) directly select a framework to make decisions, 2) predict which framework the people around it will use to make decisions and 3) predict a framework that helps the robot explain its decision making.

The data described in this paper are being used to develop an architecture that will allow a robot to flexibility and dynamically use different underlying ethical frameworks to address diverse moral problems. The data presented here are being used to generate a set of cases that forms the ethical database to be used by the robot to make action recommendations. High-level features that have been captured in the data will be used to directly index and select a case if a close match exists or probabilistically select a case based on a distance metric if a good match is unavailable. The index features for a case include risk measures and emotional models such as frustration that arbitrate among the cases provided within a given ethical framework. We are currently implementing this system and intend to test it soon.

There are several important limitations to this study. First, self-reports of one's expected behavior when faced with an ethical dilemma can differ from actual behavior when the situation arises [22]. Hence, we can only speculate as to how our subjects would actually behave if presented with these situations. Another limitation of this research is that several experts noted in their comments that additional context is needed to make an informed decision regarding how an ethical framework relates to a scenario decision. As mentioned previously, we chose to limit the scenario context to the features we were interested. Future work could explore how expert opinions might change if more information is provided, but it is unclear what or how much information would be needed to satisfy the expert's request. There may also be a limitation in terms of the generalizability of results given the relatively small sample sizes for both of the subject populations and that only participants from the United States were recruited for the surveys. Finally, as one would expect, the expert's opinions related to if and how different ethical frameworks could be applied to the scenarios differed greatly. Our data captures these differences in their comments. Unfortunately, the expert comments did not lend themselves to analysis due to the length of the comments and because of their reflective and, at times circular nature. Future research could address this issue by using structured interviews or related techniques to examine how each framework could be used to address the different scenarios.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1849068 and 1848974. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. L. Jiang, J.D. Hwang, C. Bhagavatula, R. L. Bras, M. Forbes, J. Borchardt, ... & Y. Choi, *arXiv preprint arXiv:2110.07574* (2021).
2. R. Noothigattu, D. Bouneffouf, N. Mattei, R. Chandra, P. Madan, K. R. Varshney, ... & F. Rossi, *IBM Journal of Res. and Dev.*, *63*(4/5), 2-1 (2019).
3. F. Rossi and N. Mattei, *Proc. of the AAAI Con. on Art. Int.* Vol. 33, No. 01, pp. 9785-9789 (2019).
4. H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, *arXiv preprint arXiv:1812.02953*. (2018).
5. D. Abel, J. MacGlashan, and M. L. Littman. *Workshops at the thirtieth AAAI conf. on artificial intelligence*. (2016).
6. M. Anderson and S. L. Anderson, AI Magazine, 28(4), 15. (2007).
7. D. Ross, The Right and the Good. Oxford: Oxford University Press, (1930).
8. P. Bello, and S. Bringsjord, Topoi, 32(2), 251-266 (2013).
9. J. A. Blass, and K. D. Forbus, In AAAI, pp. 501-507 (2015).
10. D. Abel, J. MacGlashan, and M. L. Littman, In AAAI Workshop: AI, Ethics, and Society, vol. 92 (2016).
11. S. M. Retzinger, Violent emotions: Shame and rage in marital quarrels. Sage (1991).
12. M. Anderson, and S. L. Anderson, (Eds.). Machine ethics. Cambridge University Press. (2011).
13. R. C. Arkin, and P. Ulam, IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09), Daejeon, KR (2009).
14. P. Lin, K. Abney, and G. A. Bekey, Robot ethics: the ethical and social implications of robotics. The MIT Press (2014).
15. W. Iba, and P. Langley, In Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 33, No. 33 (2011).
16. A. F. Beavers, In Association for practical and professional ethics, eighteenth annual meeting, Cincinnati, Ohio, March (pp. 5-8) (2009).

17 S. Bringsjord, K. Arkoudas, P. and Bello, Toward a General Logicist Methodology for Engineering Ethically Correct Robots. IEEE Intelligent Systems 21(4): 38– 44. (2006).

18 C. Grau, IEEE Intelligent Systems, 21(4), 52-55 (2006).

19 J. Rogers, and M. Holm, "Performance assessment of self-care skills test manual (version 3.1)," Pittsburgh, PA (1984).

20 S. Russell, D. Dewey, and M. Tegmark, AI Magazine 36, no. 4: 105-114 (2015).

21 B. Kuipers, In AAAI Workshop: AI, Ethics, and Society (2016).

22 D H. Bostyn, S. Sevenhant, and A. Roets. *Psychological science*, *29*(7), 1084-1093 (2018).

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.007

IS IT POSSIBLE TO DEVELOP THE MORALITY OF AI ROBOTS?

SUNYOUNG SONG
Department of Ethics Education, Gyeonsang National University, 501 Jinjudaro
Jinju, Gyeonsangnamdo 52828, Republic of Korea
E-mail: sunyoung_song@gnu.ac.kr
www.gnu.ac.kr

This paper aims to answer the question by exploring the development of morality in AI robots connected with moral education in human beings. In this paper I would like to explore three features of the development of morality in AI robots. First, the morality of AI robots is to be in the context of ethical lives in humanity. Second, to develop the morality of AI robots is not to solve hypothetical problems but to recognize the conditioned autonomy in the end of the development of morality. Third, the future task to develop the morality of AI robots is to apply ethical algorithms and reasonings to moral machines.

1. Introduction

This paper aims to answer the question by exploring the development of morality in AI robots connected with moral education in human beings. In MIT's Moral Machine, for example, various moral decisions of self-driving cars by humans have been collected as a design of a moral platform. It is evitable, of course, to make any decision in a moral dilemma that enables us to do wrong or right things without any postponement. our decisions may be moral even in routines connected with social relations. The platform of MIT's Moral Machine has imagined that 'I have to kill others', considering their ages, species, social positions, etc. The possibility that a self-driving car can kill humans is the nature of the car as a machine as well as that it can give us a sort of convenience. While the situation where the car kills humans is imaginary, our decision to kill someone is the source of morality in the car as data on how to control it.

In this paper I would like to explore three features of the development of morality in AI robots. First, the morality of AI robots is to be in the context of ethical lives in humanity. Second, to develop the morality of AI robots is not to solve hypothetical problems but to recognize the conditioned autonomy in the end of the development of morality. Third, the future task to develop the morality of AI robots is to apply ethical algorithms and reasonings to moral machines.

2. The Foundation of the Morality of AI Robots

In this chapter, the main concern is the source of the morality of AI robots. There are two questions we have to consider: 'what kind of morality can AI robots learn? and what kind of morality should we make AI robots learn?' At a glance, the common phrase in those questions is 'what kind of morality', but different from the subject of morality. The former is that AI robots learn morality for themselves in the process of programming, the latter is that the morality of AI robots is directly controlled by humans. In this chapter I would pay attention to the former.

The topic suggested in the first question can be ethical as well as technical without the consciousness of AI robots, because of the nature of AI. The conceptions of artificial intelligence are different in various fields. In a study of the definition of AI in UNESCO, various definitions of AI are related to different disciplinary approaches such as computer science, electrical engineering, robotics, psychology, or philosophy. According to UNESCO, there are two general aspects: one usually labeled as 'theoretical' or 'scientific' and the other one as 'pragmatic' or 'technological'. In terms of each aspect, the main questions of 'theoretical' or ‘scientific’ AI are like ‘what is meant by intelligence and how to distinguish natural from artificial intelligence?’; and the aspect of ‘pragmatic’ or ‘technological’ AI is applied to many areas of modern life, such as transport, medicine, communication, education, science, finance, law, military, marketing, customer services or entertainment. (UNESCO, 2019:5-6).

It is very difficult, however, to give a definite explanation on those aspects of AI, because any questions of AI are not separated in various areas of human lives. For example, it is common to see an AI system that can analyze users’ health data in their smartphones. Technically, a specific data of a user's health category such as walk, heart rate, pulse and weight is collected and utilized by programming in the system. Physically and individually, it is very helpful for the weak or the patient of heart disease to check his or her health by using smartphones in daily life. It is the exact data-based AI system that makes us enjoy the complicated benefits of smartphones.

At that moment, the one thing to be neglected is the subjectivity of the categorized data and its heuristics in the contexts of human lives. In the conception of big data, a normal range of safe and healthy pulse has been generalized, so that AI systems can see the user as being normal or unusual without any judgment of a heart specialist. However, there is a problem of generalized range that makes it possible to regard someone who is normal as abnormal. In the same vein, then, another AI system can recognize him or her as a ‘criminal’ based on big data of generalized criminals. In fact, many institutions such as the national police or FBI have got more exact information and arrested specific criminals. In the level of high technology, huge data will make it unnecessary for the agency to investigate many people for days. To be extended, the generalization of AI systems that humans can create and produce has been from a variety of data in everyday life. AI systems are operated by a given algorithm to solve particular problems and achieve some performances and, at the same time, are contextualized in the network of big data that interacts with our daily life.

Now this feature can give us a crucial clue of the morality of AI robots. Of course, AI cannot get any notion of morality. To make and utilize AI robots in our lives, nevertheless, has been still involved with ethical issues. By learning data and interacting with humans, AI robots of life have acted as agents in various fields. They can inspire us with a kind of beauty that they create by learning and interpreting human arts or be a teacher as well as a friend in communication with the weak and the old in the social welfare system. In that sense, AI robots have had the mind, knowledge, thinking, action, and memory of humans by utilizing our data, not creating new minds that we cannot recognize.

In ethical view, thus, the morality of AI robots is based on data digitized in the context of ethical lives. As seen above, it is impossible that AI robots have their own morality, but we are fearful of their moral impact. AI robots cannot have any moral intentions in their works. However, the conception of the morality of AI robots seems to us that they have a certain influence on interaction with humans, for example, social care robots that interact with patients, the old and the weak. The specific purpose of AI robots, strictly speaking, has embodied programs and orders for humans who need them. The autonomous and technical ability to learn and run the programs by utilizing data of human lives has made it possible to recognize AI robots as humans. In the digitized world the notion of data is seen as various parts of huge narration of historical and ethical lives.

3. The Conditioned Autonomy of AI Robots

One of clear contradictions in using terms of AI robots and systems is the conception of AI’s ‘autonomy’. In the engineering-oriented view, the word ‘autonomy’ of AI is to learn data, analyze given environments, and carry out any program without any direct command of humans. Without some functions like learning and analyzing, its work is only automatic by machine, but autonomous in the system such as deep learning. In an ethical view, however, the conception of autonomy is limited to human beings. As seen in Kantian view, for example, a moral agent of good will act autonomously in the world of universal moral law: to confirm our subjective criteria with universal law and to treat someone as an end. In a strict sense, it is impossible to use the usage of the autonomous AI robots in the view that only humans have free and good will, that is, the notion of autonomy.

The autonomous learning of AI robots is one of the important functions in designing and making them. In that sense, it is obvious that any judgments and doings of AI robots are not creative. The feature that AI robots learn and suggest a sort of solutions to human problems without man's help is conditioned in the network of human relations. It is the fundamental goal to develop the morality of AI robots. Moral rules or orders that AI robots learn and do autonomously are basically perfect only in the agreement or consensus between stakeholders. Therefore, the final point that AI robots should recognize in autonomous learning systems is to find out that they for themselves can never calculate any universalized solutions on partiality in human relation. That is, the autonomy of AI robots is conditioned in the network of human relations.

In the level of conditioned autonomy in AI robots, then, it seems a serious question: is it to give rise to moral and cultural relativism? A sort of specialized and contextualized AI robot is very useful to specific environments but can do opposite judgment in different situations and data. However, this question is only proper if the AI robot has the independent autonomy of humans. In this question, thus, the nature of the conditioned autonomous robot emerges from diversity and contextualization.

4. Conclusion: Future Task to Develop Ethical Algorithm

It would be concluded that the future task is to develop and apply ethical algorithms and reasoning to moral machines, even though there will be various and huge innovations in AI robots and systems. As a useful tool to enhance the moral capacity of humans, AI robots can be developed in various human parts. Concerning the future of AI robots, it seems to be a scenario in SF movies whether or not AI robots do autonomously recognize and replace the biological and spiritual functions of humans. But it is also obvious that AI robots are controlled by technology. The essential task is to consider and develop ethical algorithms and reasonings to moral machines to apply them in practices. Due to the variety of humans, differentiated moral standards and programs of AI robots may justify a sort of moral relativism. That is immediately against the nature of the morality of AI robots, so to speak, 'conditioned autonomy' as a useful tool to enhance moral capacity. AI robots are never beyond the network of human relations and ethical lives, which means that the various and different moral standards of AI robots are relational, not relative.

In this draft, for example, I would like to outline two schemes related to ethical views in ethics education. First, in moral development of L. Kohlberg, AI robots and systems is to learn data and analyze criteria of right and wrong in ethical lives with hypothetical and practical dilemma of students. The main purpose is to draw out the competence of moral development and constitute their morality for themselves. Through ethical issues, various data that students suggest is a kind of ethical big data as well as a set of narration of individual life. An algorithm of AI robots to develop morality of justice in Kohlberg's view, therefore, is a conditioned interlocutor that enables human students to put themselves into others' shoes and learns and analyze thoughts and criteria of right and wrong in students and gives their teachers feedback on individual data (Sunyoung Song, 2021b: 44-45). Second, it is considerable to apply ethical theories to various situations of ethical algorithms. A principle of avoiding and removing 'pain' based on J. Bentham's Deontology is an instance in this draft. As seen in the investigation of the Springs of Action in Bentham, an ethical algorithm to eliminate pain can proceed seven stages to do practices (Sunyoung Song, 2021a: 176).

Stage 1. Premise: Pain is evil. The word 'pain' is private as well as public, not private, public or selfish.

Stage 2. To check the motive of actions. Of right or wrong, the motive of an act is to calculate its results.

Stage 3. To extend the category of interests. There are three categories: self-regarded, self and others-regarded, others-regarded.

Stage 4. To classify the sorts of pain. The sorts of pain that Bentham suggests are 14.

Stage 5. To evaluate pain. This is the calculation of Bentham to maximize general happiness.

Stage 6. To motivate the calculation. to do practice of removing pain in the final step needs to confirm the motive of the decision: praised or criticized.

Stage 7. To do practice and give feedback. It is important to do practice of the motive and the calculation of removing pains.

References

1. Sunyoung Song, A Study of Teaching Design of AI Ethic Education based on J. Bentham's Deontology and Ethical Algorithm to eliminate Pains: with emphasis on Revisions and Contents of National Curriculum in Moral Education and Informatics. *PHILOSOPHY·THOUGHT·CULTURE*, vol. 39, 18-38 (2022). DOI: http://dx.doi.org/10.33639/ptc.2022.39.002
2. Sunyoung Song, A Study of the Education of Moral Development in J. Bentham's Deontology: with emphasis on the probability of ethical algorithm to utilize AI robots. *Journal of Public Society*, vol. 11, no. 2, 150-180 (2021a). DOI: http://dx.doi.org/10.21286/jps.2021.05.11.2.150
3. Sunyoung Song, How Are AI Robots Reversible? with emphasis on the Application of AI Robots to the Theory of Moral Development in L. Kohlberg as Teaching Tools. *Journal of Ethics Education Studies*, vol. 64, 27-54 (2021b). DOI: http://dx.doi.org/10.18850/JEES.2021.61.02
4. Sunyoung Song, The Possibility of Narratives and Empathy between Medical Care Robots and Human Patients. *Human Beings, Environment and Their Future* (HEF), no. 18, pp. 49-76 (2017). DOI: http://dx.doi.org/10.34162/hefins.2017.18.003
5. Sunyoung Song, Bioethics, Moral Education, and the Potential Relationship between Humans and robotics in the future: utilizing the motives of Avatar and Surrogates. *Journal of English Cultural Studies*, vol. 9, no. 2, pp. 23-48 (2016). DOI: http://dx.doi.org/10.15732/jecs9.2.201608.23
6. UNESCO, Preliminary Study on the Ethics of Artificial Intelligence (2019).

SECTION–3

AUTONOMY AND AUTONOMOUS SYSTEMS

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.004

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

CATHERINE MENON and SILVIO CARTA

University of Hertfordshire, Hatfield, England, United Kingdom

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

www.herts.ac.uk

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

1. Introduction

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

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

Step 1: Provide AS description

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

Step 2: Identify relevant ethical principles

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

- System promotes human physical safety
- System obeys human commands

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

Step 3: scenario construction

Teams are then asked to generate scenarios for this system in which two of the selected ethical principles are in conflict with each other. For example, a team may propose the scenario where a user commands their assistive robot not to remind them about medication, creating a conflict between the ethical principles of "system promotes human physical safety" and "system obeys human commands". A set of guide words based on Hazard and Operability Analysis processes will be provided to the teams to assist in brainstorming these scenarios.

Step 4: Identify constraints

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

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

Use of a design tool

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

Recording outcomes

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

3. Conclusions and next steps

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

Figure 1. The EETAS-TOY gamified tool

References

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

[a] www.weandai.org

ICRES 2022: 7th International Conference on Robot Ethics
and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.008

SENTRY ROBOTS IN ACTION: ETHICAL AND LEGAL ISSUES OF AUTOMATED WEAPON IN SOUTH KOREA

HAN-KYUN KIM
Korean Institute of Criminology & Justice, 114 Taebong-No, Seocho-Gu, Seoul, Korea
E-mail: hankyun@kic.re.kr

South Korea is one of the leading countries both in the development of robot industry and the deployment of lethal robot weapons. The development of the SGR-A1 has brought some ethical controversy regarding the issue of autonomous killing and collateral damage. Based on the ethical and legal policy consideration over the issue of robots for military purposes, which must be the most serious or imminent threat of AI robots, the Paper examines how to develop ethics-based, future-proof law and policy on the development of robots.

1. Officially Unknown Rise of the Automatic Weapon – A South Korean Case

1.1. *World's first-ever 'killer robots' in action?*

South Korea marks one's of the highest level of industrial robots operating in factories around the world today,[a] may also be the world's first country where Lethal Automatic Weapons System (LAWs) in action.

It had been reported that a gun-toting sentry robot, developed by Samsung Techwin Co. for the South Korean government is dispatched in the De-Militarized Zone (DMZ) between the two Koreas. SGR-A1 robot [see Figure 1] uses a low-light -camera and pattern recognition software to distinguish humans from animals or other objects and, if necessary, can fire its built-in machine gun.[b]

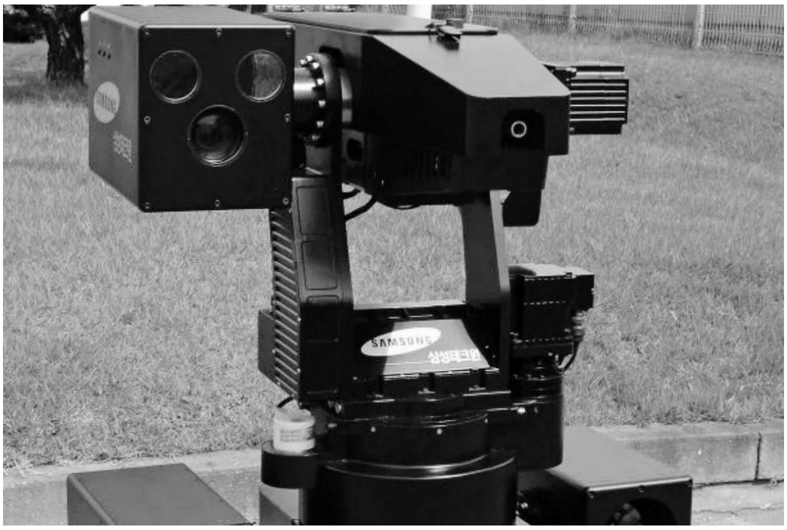


Figure 1. A weapons-grade sentry robot from South Korea[c]

[a] International Federation of Robotics, World Robotics Report 2020(2020)

[b] Jean Kumaga, *A Robotic Sentry for Korea's Demilitarized Zone* (https://spectrum.ieee.org/a-robotic-sentry-for-koreas-demilitarized-zone)

[c] South Korea Develops Machine-Gun Sentry Robot (Posted on 05/10/2006, https://www.robotshop.com/community/blog/show/south-korea-develops-machine-gun-sentry-robot)

South Korean defense ministry sees its goals as to transform the current guard and observation mission on fronts conducted by soldiers into a robot system. Such system may have some advantages, as they do not require communication links, fewer soldiers are needed, and because they are not humans, war crimes such as rape and indiscriminate killing can be reduced.

1.2. *Controversies over the robot in military action*

The development of the SGR-A1 has brought some controversy regarding the issue on ethics of autonomous killing, and innocent killings through collateral damage; The SGR-A1 presumes any person entering the DMZ is an enemy and, upon detection, will attempt to identify the target through voice recognition. If a proper access code is not provided within a short amount of time, the system can choose between sounding an alarm, firing rubber bullets, or engaging the target with other weapons. The system can also be overridden by an operator, who can also communicate via built-in microphone and audio system.

It is believed that human-in-the-loop system is applied to the SGR-A1, which means that the weapon must wait for commands from a human operator before acting upon its targets. However, Committee on International Security and Arms Control and Human Rights Watch believe the SGR-A1 has human over the loop capabilities and would ultimately increase civilian casualties.

2. Ethical and Policy Consideration over the issue of Robots for Military Purposes

2.1. *Ethical concerns against the threats from any automated weapons*

LAWS are unethical because humans are no longer involved in the act of killing, which absents morality from warfare by ultimately pitting a machine against a human. LAWS increase the risk of conflict and war by shifting risks from the nation's soldiers to machines, which distances civilians from war and thus changes how the military and government think about going to war. LAWS introduce a difficulty of assigning responsibility and accountability if scenarios including civilian casualties and violation of international law arise.

Since 2012, the Campaign to Stop Killer Robots has been working to ban fully autonomous weapons. There has been increasing support for such a ban. In 2015, more than 3,000 experts in artificial intelligence and robotics research signed an open letter calling for a treaty to ban lethal autonomous weapons. In 2019, UN Secretary-General also called for banning fully autonomous weapons, saying “machines with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant and should be prohibited by international law.”[d]

United States, China, Israel, South Korea, Russia, and the UK are all developing weapons systems with significant autonomy in the critical functions of selecting and attacking targets. This represents a dangerous new arms race for autonomous weapons. As of 2021, South Korea has not endorsed the call for prohibiting fully autonomous weapons.

In 2019, the South Korean government acknowledged the need for “a concerted response against … potentially grave challenges” posed by lethal autonomous weapons systems and said that

[d] Machines Capable of Taking Lives without Human Involvement Are Unacceptable, Secretary-General Tells Experts on Autonomous Weapons Systems (https://www.un.org/press/en/2019/sgsm19512.doc.htm)

“Korea is … committed to joining international efforts in the long-term to develop realistic international norms” regarding their use.

2.2. *Ethics and robots at issues in South Korea*

In 2018, more than 50 international scholars signed a letter calling for a boycott of Korea Advanced Institute of Science and Technology (KAIST) over concerns that it was developing autonomous weapons. On the other hand, some AI Charter of Ethics have been created: Draft of the Robot Ethics Charter by the Ministry of Commerce Industry and Energy in 2007, Ethical Guidelines for Intelligence Information Society by the Ministry of Science and ICT & National Information Society Agency in 2018, Intelligent Government Ethics Guideline for Utilizing Artificial Intelligence by Korean Internet Ethics Association & National Information Society Agency in 2018, and the Charter of Artificial Intelligence Ethics by Korea Artificial Intelligence Ethics Association in 2019. The core principles of such guidance are this: humans should be able to control AI, and the purpose of AI should be limited so that it cannot be used to injure or kill humans.

South Korean government focuses on the realization of people-centered AI and human-like AI we establish a global level of AI code of ethics based on social debate and consensus. To prevent the dysfunction of AI, It proposes the establishment of a quality management system that verifies reliability and safety in response to the proliferation of AI products and services . It also prepares to establish AI ethical standards that are consistent with global norms by identifying and analyzing the AI Code of Ethics and discussion trends in international organizations and major countries and setting up action plans. [e]

3. How to develop ethics-based, future-proof law and policy on the development of robots

3.1. *When a robot be threat to human, what is to be done with ethics and law?*

While South Korea has made some efforts to appropriately response to the threat of lethal robot weapons, South Korea has been developing a killer AI robot. This may be because provisions involving lethal AI will be difficult to implement because there is already competition for the development of AI weapons among various international military powers. So the practical task will be to devise some AI-risk management system, legal framework of responsibility, and culture of social communication over the AI threat issues, that is national AI governance system which covers possible threats of and from robot in action.

3.2. *On its own ‘Trustworthiness’ of the robots as automated weapons*

Intelligent Robots Development and Distribution Promotion Act of 2008 which amended in 2016, aims to contribute to enhancing the quality of life of citizens and the national economy by establishing and promoting a policy on the sustainable development of the intelligent robot industry to facilitate the development and distribution of intelligent robots. Article 2, para.2 of

[e] Government of Republic of Korea, National Strategy for Artificial Intelligence (2019) pp.47-49.

the Act defines "intelligent robot ethics" as a code of conduct established for persons involved in the development, manufacture, and use of intelligent robots in order to prevent various kinds of harmful or adverse effects, such as destruction of social order, that may arise from the development of functions and intelligence of intelligent robots, and to ensure intelligent robots contribute to enhancing the quality of life of human beings.

For an intelligent and safe robot in use, the governance system and legal framework has to be built upon the ethical value of 'trustworthiness', so that lethal weapon should be lawful, complying with all applicable laws and regulations; it should be ethical, ensuring adherence to ethical principles and values; and it should be robust, both from a technical and social perspective.

References

1. Jean, Kumagai, *A Robotic Sentry for Korea's Demilitarized Zone*, IEEE Spectrum 44(3) (2007)
2. Jon, Rabiroff, *Machine Gun Toting Robots Deployed On DMZ*, Stars and Stripes (2010)
3. Ross, Anderson, *Cyber and Drone Attacks May Change Warfare More Than the Machine Gun*, The Atlantic (2012)
4. Lee, Yeon Hee & Byun, Sung Yong, *An Ethical Consideration on Killer Robots*, The Journal of Korea Elementary Education Vol. 31(2020)
5. Korea Peace Now, *Factsheet: Killer Robots and South Korea* (2020)
6. Ha Hwang & Min-Hye Park, *The Threat of AI and Our Response: The AI Charter of Ethics in South Korea*, Asian Journal of Innovation and Policy (2020)
7. Jung, Chea Yun, *Building an Ethical Intelligent Robot in the Intelligent Information Society*, East Asian Law Review 18(1) (2018)
8. Kim, Gunoo, *Robot Ethics vs. Robot Law: Apart and Together*, Studies on Legal Philosophy 20(2)(2017)
9. Government of Republic of Korea, *National Strategy for Artificial Intelligence* (2019)

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.009

A Framework for Testimony-Infused Automated Adjudicative Dynamic Multi-Agent Reasoning in Ethically Charged Scenarios

Brandon Rozek and Michael Giancola and Selmer Bringsjord and Naveen Sundar Govindarajulu

Rensselaer AI & Reasoning (RAIR) Lab, Rensselaer Polytechnic Institute (RPI),
110 Eighth Street, Troy, NY 12180, USA
E-mail: {*Rozek.Brandon, Mike.J.Giancola, Selmer.Bringsjord, Naveen.Sundar.G*}*@gmail.com*

In "high stakes" multi-agent decision-making under uncertainty, testimonial evidence flows from "witness" agents to "adjudicator" agents, where the latter must rationally fix belief and knowledge, and act accordingly. The testimonies provided may be incomplete or even deceptive, and in many domains are offered in a context that includes other kinds of evidence, some of which may be incompatible with these testimonies. Therefore, before believing a testimony and on that basis moving forward, the adjudicator must systematically reason to suitable *strength* of belief, in a manner that takes account of said context, and globally judges the core issue at hand. To further complicate matters, since the relevant information perceived by the adjudicator changes over time, adjudication is a nonmonontonic/defeasible affair: adjudicators must dynamically strengthen, weaken, defeat, and reinstate belief and knowledge. Toward the engineering of artificial agents capable of handling these representation-and-reasoning demands arising from testimonial evidence in multi-agent decision-making, we explore herein extensions to one of our prior *cognitive calculi*: the *Inductive Cognitive Event Calculus* ($\mathcal{IDCEC}$). We ground these extensions in a recent, tragic drone-strike scenario that unfolded in Kabul, Afghanistan, in the hope that use by humans of our brand of logic-based AI in future such scenarios will save human lives.

Keywords: testimony-infused decision-making; multi-agent reasoning; argument adjudication

1. Introduction

Human agents often believe various propositions because they perceive part of their environment. Such an agent $\mathfrak{a}_h$ often believes for instance that there is a cup of coffee on the kitchen table because it sees the cup there courtesy of its own unaided sensors (eyes, e.g.); and in this case, all things being equal (e.g., the perceiver is not severely intoxicated), $\mathfrak{a}_h$ now as a result believes that there is a cup of coffee on the table. This basic picture stands at the heart of at least logic-based (= logicist) AI ([1], esp. Chap. 7 "Logical Agents;" and see as well dedicated treatments of logicist AI, e.g. [2]) and cognitive robotics of an overtly logicist sort [3], and is also a part of the very foundation of the empirical study of human cognition in information-processing terms (see e.g. [4]). However, agents, whether human or (present-day) artificial, are, we can all agree, not omnipresent; for this reason they often rely upon other agents to exceed the range of their own unaided sensors, by taking from these others *testimonies* (a term we use in its general sense, not in any narrow legal sense). A human agent located outside the kitchen may call to another agent inside it, "Is my coffee on the table in there?", and if hearing back an affirmative may rationally believe as a result that there is a cup of coffee on the kitchen table.

We shall in the present paper take a testimony to essentially have the basic shape of a triple $(\mathfrak{a}_w, \psi, \mathfrak{a}_{adj})$, where ψ is a declarative formula shared (via some form of communication, which make use of natural language expressing ψ) by a witness agent $\mathfrak{a}_w$ to an "adjudicator" agent $\mathfrak{a}_{adj}$. Adjudication is needed because whether it's rational for $\mathfrak{a}_{adj}$ to believe ψ at a given time frequently hinges on myriad factors, including competing, incompatible ones; and some of these competing factors can be testimonies themselves. The adjudicator in the case of the coffee example may receive in addition to an affirmative in response to a query, a negative one — and now what should the adjudicator believe about the availability of desired caffeine? Realistically, we cannot assume that witness agents presenting testimonies are faultless. Such agents may have compromised perception or even ulterior motives. Therefore, when collecting and forming beliefs from testimonies, the adjudicator must reason over relevant, available information before fixing belief. And of course rational belief fixation engaged through time as the world changes and offers up new information, as has long been known in AI, is a temporally extended reasoning process that can't be exclusively deductive: this must pass into the realm of *inductive* logic, where inference is non-deductive, and uncertainty measures of some sort are used. In particular, new testimonies at time t may strengthen, weaken, or even defeat each other at that time, and may do the same to testimonies issued prior to t. Hence from the standpoint of logic-based (= logicist) AI and cognitive science, adept defeasible/nonmonotonic reasoning must be part of the adjudicator's cognitive arsenal.

To ground the rather admittedly abstract concepts and structures sketched in the previous paragraph, and the logico-mathematics behind them (which, in the form we prefer, we share soon: §3), we shall rely below upon an illuminating (and certainly sobering) case study of a recent drone strike in Kabul, Afghanistan. At the end of August 2021, as is widely known, US forces were evacuating Afghanistan. Three days before the incident we soon study, an ISIS K suicide attack killed 13 US troops and more than 60 afghan civilians [5]. The desire to prevent another attack was understandably high, as were tensions. In this emotionally and ethically charged context, authorization was given to employ kinetic counter-measures even under uncertainty, and as a matter of fact, such authorization was used — with tragic loss of innocent life. To use AI (or at least our brand of it) to prevent such tragedies in the future, automated reasoners must support, through time, ethical reasoning and counter-reasoning. We specifically need, as well, automated reasoners with the capability to detect and resolve inconsistencies arising from competing testimonies, arguments, and positions on profound moral matters. But this is only one desideratum (d_2, as will be shortly seen) among seven that constitute the requirements for the kind of capability our automated reasoning must have.

Enough introduction. The plan for the remainder is as follows: In the next section (§2), we enumerate, with brief exposition, the desiderata just alluded to. What follows next (§3) is a summary of the formal framework used in the work we report herein. Following that (§4), we return to the case study sketched in the previous paragraph, and establish (familiar, and broadly if not universally affirmed) conditions needed to permit a strike under the relevant type of contextual conditions by U.S. forces. We discuss related work and alternative approaches in §5, and compare them against our desiderata for automated adjudication of key information in the case study. In §6, we discuss the testimonies from outside intelligence sources and explain how, at least in our view, a rational defeasible system should handle them. Relevant and cognitively plausible inductive arguments are provided and treated in §7, along with a demonstration of our automated reasoner and how it can infer and adjudicate beliefs in a time-feasible manner that at least suggests the viability of AI-infused multi-agent decision-making in future situations analogous to the case study. We then close out the paper (§9) with with some final remarks, and recommended future actions.

2. Reasoning-System Requirements

Needless to say, any proposed set of requirements, or desiderata, for an automated reasoner (or for an ensemble of such systems) will directly reflect the general objectives and methodological orientation of the researchers and engineers involved in the pursuit at hand. We do not pretend that our overarching objectives are universally affirmed. For instance, for us, any formal computational logic that fails to formalize and enable sophisticated *intensional* reasoning is unacceptable (relative to the applications we tend to emphasize), for reasons going back to the Frege, who while giving us the first fully rigorous and top-to-bottom presentation of first-order logic = $\mathscr{L}_1$, also presented us with the challenging observation that some rational agent $\mathfrak{a}$ can have beliefs about the morning star s_m, and radically different beliefs about the evening star s_e, and have no clue whatsoever that — expressed in the terms of extensional $\mathscr{L}_1$ — $s_m = s_e$.[a] This is specifically desideratum d_6 in the set of such, one we dub '$\mathcal{D}$', which is that ... An automated reasoner of the kind we seek must:

Desiderata ($\mathcal{D}$)

- d_1 be defeasible (and hence — to use the term frequently employed in AI — nonmonotonic) in nature through time;
- d_2 be able to resolve inconsistencies (of various sorts, ranging e.g. across ω-inconsistency to "cognitive inconsistency" (e.g. an agent $\mathfrak{a}$ believing both ϕ and $\neg\phi$) to standard inconsistency in bivalent extensional logic) when appropriate, and tolerate them when necessary in a manner that fully permits reasoning to continue;
- d_3 make use of values beyond standard bivalence and standard trivalence (e.g. beyond the Kleenean TRUE, FALSE, UNKNOWN trio), specifically probabilities *and* likelihood values or strength-factors (the latter case giving rise to multi-valued inductive logics corresponding to the cognitive calculus $\mathcal{IDCEC}$ used below);
- d_4 be argument-based, where the arguments have internal inference-to-inference structure both in terms of declarative formulae (and possibly diagrams) and inference schemata (as opposed to purely abstract, meta-logical formalisms such as those of [7]), so that detailed step-by-step verification is possible, and over justification/explanation is available;
- d_5 have specified inference schemata (which sanction the inference-to-inference structure referred to in d_4), whether deductive or inductive, that are machine-checkable;
- d_6 be able to allow automated reasoning over the socio-cognitive elements of knowledge, belief, desire, perception, communication, emotion etc. of relevant artificial and human agents, where these elements are irreducibly intensional;
- d_7 be able to allow automated reasoning that can tackle Turing-unsolvable reasoning problems (in, of course, particular instances), e.g. queries about provability at and even above the *Entscheidungsproblem* (e.g. at and above Σ_1^0 and Σ_1^1 in the Arithmetical and Analytical Hierarchies, resp.).

3. Formal Background

We make use herein of our previously erected formal framework for logicist AI and specifically automated reasoning, the chief component of which is a *cognitive calculus* $\mathcal{C}$ within an uncountably infinite family $\mathscr{C}$ of such. Full coverage of the family of cognitive calculi is out of scope.[b] We rely rather heavily in the present paper on the exemplar cognitive calculus

[a] A nice overview of intensional logic is given in [6], which in fact does discuss Frege and the example of Venus.

[b] Very briefly, the first building block of a cognitive calculus is simply a purely extensional and purely deductive *logical system* defined as in standard mathematical logic (e.g. in coverage of Linström's Theorems

presented momentarily, and used thereafter in the case study, to give readers (presumed to largely be cognoscenti) a good sense of what a cognitive calculus is; this cognitive calculus is the Inductive Deontic Cognitive Event Calculus ($\mathcal{IDCEC}$), an inductive relative of the purely deductive $\mathcal{DCEC}$. For further information, the following resources among others are available: An efficient introduction to the family $\mathscr{C}$ is provided in [13]; use of a particular cognitive calculus for an ethically charged application handled by $\mathcal{DCEC}$ is provided in [14]; the first cognitive calculus that appeared in print is defined and used in [15]; for those who are more on the side of cognitive science than engineering-oriented AI, a sub-family of cognitive calculi are introduced and used in implemented form in [16], and sustained coverage in survey style of how formal logic can be used for cognitive modeling can be found in [17,18]. One final point before passing to specifics: A distinctive aspect of cognitive calculi is that they can be *heterogeneous*: their formal languages and inference schemata can allow diagrams and other pictorial elements, an approach given in the logic Vivid [19].

This works extends the $\mathcal{IDCEC}$ introduced in [20]. Briefly, a cognitive calculus consists of two parts: a signature and a set of inference schemata. The signature of the verison of $\mathcal{IDCEC}$ deployed herein is given in the box titled $\mathcal{IDCEC}$ **Signature**. It consists of three components: (1) a set of sorts (e.g. Agent, Action, etc. in order to capture states of the world and how it changes through time), (2) a set of types, and (3) a set of syntactic forms, including those of propositional and first-order logic as well as cognitive epistemic operators **B**elieves, **C**ommon-knowledge, **S**ays, and **P**erceives. The formulae are read in a fairly intuitive way. For example, $\mathbf{S}(a, b, t, \phi)$ is read as "Agent a *says* ϕ to agent b at time t."

$\mathcal{IDCEC}$ **Signature**

$$S ::= \mathsf{Agent} \mid \mathsf{ActionType} \mid \mathsf{Action} \mid \mathsf{Moment} \mid \mathsf{Fluent}$$

$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{cases} \psi : \mathsf{Formula} \mid \forall t : \phi \mid \exists t : \phi \\ \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \phi \rightarrow \psi \mid \phi \leftrightarrow \psi \\ \mathbf{B}^{\sigma}(\mathfrak{a}, t, \phi) \mid \mathbf{C}(t, \phi) \mid \mathbf{S}(\mathfrak{a}, b, t, \phi) \mid \mathbf{P}(\mathfrak{a}, t, \phi) \end{cases}$$

Cognitive uncertainty is captured through likelihood values shown in Table 1 and used

in [8]), which is then modified in three key ways: Move (i) is that all model-theoretic semantics are discarded but selectively recast and pulled back in purely proof-theoretic terms, by moving the central meta-logical expressions in such semantics (e.g. that some formula ϕ is satisfied by some interpretation $\mathcal{I}$, customarily written '$\mathcal{I} \models \phi$') into object-level syntax. And (ii), some set $\mathcal{I}$ of natural (hence when the schemata are deductive in nature, we specifically have natural-deduction schemata) inference schemata for a given calculus must be fixed, but are allowed to be inductive in nature (and hence draw from inductive logic, e.g. non-inferential *pure inductive logic* [9] or semi-formal inductive logic in the argument-centric tradition [10]) and make use of uncertainty measures (probabilities, likelihood values (which are used below), etc.), and done so in keeping with the third move (iii), which is the addition of modal operators that represent one or more cognitive verbs at the human level standardly covered in human-level cognitive psychology (e.g. see any standard, comprehensive textbook on human-level cognitive psychology, such as [11,12]), and regarded to be so-called "propositional attitudes" that give rise to propositional-attitude-reporting sentences, where these sentences are represented by operator-infused formulae in a cognitive calculus. Such verbs include: *knowing*, *believing*, *deciding*, *perceiving*, *communicating*, *desiring*, and *feeling* X where 'X' denotes some emotional state (e.g. possible $X = sad$, and so on. (Thus the reason we speak of a *cognitive* calculus should be plain.) "Off-the-shelf" modal logics are rejected because not only are they model-theoretic, possible-worlds-based, etc. instead of being purely inferential, but such semantics constrain what can be represented and automatically reasoned about, since e.g. perfectly meaningful English sentences are beyond the reach of any off-the-shelf logic (such as some version of quantified S5), e.g. "Selmer ought to bring it about that Brandon believes that at least three friends of Mike's know that Selmer just said 'None of Mike's friends save for one are angry'."

to ascribe a *quality* to the level of belief. This is represented formulaically as a superscript within the **B** operator and denoted by 'σ.' For example, $\mathbf{B}^4(a, t, \phi)$ can be read as "Agent a believes it is BEYOND REASONABLE DOUBT (cognitive likelihood value 4) that formula ϕ holds at time t."

As said above, we follow a proof-theoretic (or more accurately, an argument-theoretic[c]) as opposed to a model-theoretic (or, for the modal case, possible-worlds) approach. Proof-theoretic semantics for extensional logics, which avoid completely any Tarskian notion as a domain of discourse over which for example quantification ranges, was introduced by [21], extended in e.g. [22], and for a contemporary non-technical overview readers can consult [23]. The proof-theoretic semantics for cognitive calculi are beyond the scope of the present paper. An introduction to the core idea of extending proof-theoretic semantics for extensional logics to intensional ones, can be found in [24], which builds upon the natural-language-specific aspect of proof-theoretic semantics as set out in [25].

To summarize, the semantics of cognitive calculi are given exclusively via inference schemata, which dictate how new formulae can be derived and proofs can be constructed. The set of inference schemata for the version of $\mathcal{IDCEC}$ used herein is given in the box titled $\mathcal{IDCEC}$ **Inference Schemata**. Generally, an inference schema can be understood to say, "If the set of formulae above the line are true, then the formula below the line can be inferred." We will next describe how to interpret $[I_{WLP}]$, as it should then be clear how to interpret the others.

$[I_{WLP}]$ essentially implements the *Weakest Link Principle*, which dictates that an argument is only as strong as its weakest link. More formally, the schema says that if an agent a holds a set (of size m) of beliefs in formulae ϕ_1 to ϕ_m at likelihoods σ_1 to σ_m, and ϕ is provable from the set $\{\phi_1, \ldots, \phi_m\}$ *but* the set is not inconsistent (i.e. it cannot prove a contradiction), then at a later time t, a can infer a belief in ϕ, but only at the *minimum* likelihood of the beliefs used in the inference.

Given some background knowledge Γ, we desire our automated reasoner to make *ethical* decisions according to some ethical principle ρ. In $\mathcal{IDCEC}$, this will tell us whether the performance of an action α is ethically permissible, obligatory, or forbidden at time t. In section 4.1, we summarize the so-called Doctrine of Double Effect ($\mathcal{DDE}$), an ethical principle that draws from consequentialist, deontological, and divine-command ethical theories/traditions. As should be clear, nothing in the formalisms and technology that constitute the framework of our work is wed to $\mathcal{DDE}$: any credible and formalizable ethical principle can be used. (That said, $\mathcal{DDE}$ is in fact the basis for "just war" in the US and NATO.)

$\mathcal{IDCEC}$ **Inference Schemata**

$$\frac{\mathbf{B}^{\sigma}(\mathfrak{a}, t_1, \Gamma), \quad \Gamma \vdash \phi, \quad t_1 \leq t_2}{\mathbf{B}^{\sigma}(\mathfrak{a}, t_2, \phi)} \; [I_{\mathbf{B}}] \qquad \frac{\mathbf{P}(\mathfrak{a}, t_1, \phi), \quad \Gamma \vdash t_1 < t_2}{\mathbf{B}^{5}(\mathfrak{a}, t_2, \phi)} \; [I_{\mathbf{P}}]$$

$$\frac{\mathbf{B}^{\sigma_1}(\mathfrak{a}, t_1, \phi_1), \ldots, \mathbf{B}^{\sigma_m}(\mathfrak{a}, t_m, \phi_m), \{\phi_1, \ldots, \phi_m\} \vdash \phi, \{\phi_1, \ldots, \phi_m\} \nvdash \bot, \Gamma \vdash t_i < t}{\mathbf{B}^{min(\sigma_1, \ldots, \sigma_m)}(\mathfrak{a}, t, \phi)} \; [I_{WLP}]$$

[c]We note that argument-theoretic semantics are, in essence, the same as proof-theoretic semantics with one key distinction: it comes down to what differentiates a proof from a (formal) argument. Whereas a proof must be completely deductive, an argument can contain inductive/uncertain elements (e.g. uncertain beliefs, and/or inductive inference schemata), and hence an argument can conclude with a formula which contains a degree of uncertainty.

Table 1: The 13 Cognitive Likelihood Values

Numerical	**Linguistic**
6	*Certain*
5	*Evident*
4	*Beyond reasonable doubt*
3	*Very likely*
2	*Likely*
1	*More likely than not*
0	*Counterbalanced*
-1	*More unlikely than not*
-2	*Unlikely*
-3	*Very unlikely*
-4	*Beyond reasonable belief*
-5	*Evidently not*
-6	*Certainly not*

4. Case Study Part I

At the end of August 2021, the US was pulling its troops out of Afghanistan, primarily by way of the airport in Kabul. On August 26th 2021, two suicide bombers and gunmen attacked Kabul's airport, killing at least 13 US troops and 60 Afghans [5]. The desire to seek out and prevent future attacks on the airport was (naturally) acute. Six MQ-9 Reaper drones were deployed in order to search for potential ISIS K collaborators with the motive, means, and intentions[d] to bomb the airport. Once a suspect is located, a drone will track their activities to gather additional evidence, if possible. If a target is subsequently found, a sufficient level of evidence is sought, as are satisfaction of the ethical conditions. If conditions are met, then in this context, by U.S. policy, a hellfire missile may be fired to act as a counter-attack to prevent an attack on the airport.

4.1. *Conditions for Strike*

The DoD Law of War manual includes principles such as military necessity, proportionality, and distinction [28]. Military necessity justifies non-prohibited measures to end the war as quickly as possible. Proportionality limits the actions taken so that they are not unreasonable or excessive according to some utility function γ. Distinction ensures the protection of non-combatants and their objects. All of this is directly reflective of the longstanding Occidental tradition of "just war," going back to the Doctrine of Double Effect ($\mathcal{DDE}$), which originates with Augustine and was substantively refined by Aquinas. Drawing from consequentialism and deontology, $\mathcal{DDE}$ provides a set of conditions required to ethically permit an action α that may itself result in loss of life. (For an introduction to $\mathcal{DDE}$ in an AI context, see [14]; further analysis, formalization, and simulation with automated reasoning is provided in [29].) In the Kabul scenario, the ethical permissibility of firing a hellfire missile, requires a belief, at the level of *overwhelmingly likely* that ($\sigma_i \geq 4$):

$$C_1 \quad \mathbf{B}^{\sigma_1}(operator, t, \\ Capable(suspect, bomb(airport)))$$

[d] This is in general the correct approach, abstractly considered, in our opinion. We leave aside in this paper coverage of our reliance upon a formalization of the Wigmorean "MMOI" (motive, means, opportunity, and intent) pattern to persons of interest. For the background here, readers can consult [26] for a modern treatment, and [27] for original by Wigmore himself.

C_2 $\mathbf{B}^{\sigma_2}(operator, t, \mathbf{I}(suspect, t, bomb(airport)))$
$\mathcal{DDE}$ sanctions the strike.

When an operator encounters an imminent threat, a strike can be made under $\mathcal{DDE}$ as an act of *self-defense*. A large risk associated with this type of strike is the failure to detect and counter-attack within a fixed time period — which can often lead to catastrophic loss of life. In fact, an example of this is the attack on the airport three days prior, mentioned above. Therefore, if the former condition is not met, the following condition, in line with $\mathcal{DDE}$, acts as an alternative to permit a strike amid uncertainty.

C_1^* A positive belief of C_1 with $\sigma_1 > 0$.
C_2^* A positive belief of C_2 with $\sigma_2 > 0$.
C_3 $\mathbf{B}^{\sigma_3}(operator, t, \neg\exists t' > t :$ $Capable(operator, t', counterattack))$ with $\sigma_3 \geq 4$.

Conditions for C_1 We define a suspect as capable of bombing the airport if they are near the airport and are in possession of an explosive item:

$$\frac{\begin{array}{l} B^{\sigma}(operator, t, \exists item \ : \\ Near(suspect, airport) \wedge \\ Explosive(item) \wedge \\ Has(suspect, item)) \end{array}}{B^{\sigma}(operator, t, Capable(suspect, t, bomb(airport)))}$$

Conditions for C_2 Assessing the intent of an individual is of course difficult. In the present case, our approach, as announced at the outset of the paper, is to rely upon testimonial evidence; our doing so is made plain §6.

Conditions for $\mathcal{DDE}$ Below are the informal conditions which express $\mathcal{DDE}$. A formalization of $\mathcal{DDE}$ in the cognitive calculus $\mathcal{DCEC}$ can be found in [14], and expressed in a formal but meta-logical manner in [30].

$\mathcal{DDE}_1$ The action α by itself is not ethically forbidden.
$\mathcal{DDE}_2$ The net utility of α in the situation is greater than some (non-trivial) positive amount γ.
$\mathcal{DDE}_3$ The agent performing α intends only the good effects from this action.
$\mathcal{DDE}_4$ The agent does not intend any of the bad effects from α.
$\mathcal{DDE}_5$ The bad effects are not used as a means to obtain the good effects.

Returning to the scenario, the drones were equipped with a camera which the operators used throughout the day to perceive and monitor multiple video feeds, all while incorporating outside intelligence into their process of belief fixation. The statements from this scenario are derived from [31] as well as [32]. Distinction between vehicles were communicated by their make, model, and color. The one we will see repeated multiple times is a white Toyota Corolla, represented as:

$$\begin{array}{l} \forall x : WTC(x) \iff \\ \qquad White(x) \wedge Toyota(x) \wedge Corolla(x) \end{array}$$

We begin the scenario at time t_0 with an operator tracking a suspect driving a white Toyota

Table 2: Beliefs from Perception (at likelihood 5 from $I^s_{\mathbf{P}}$)

Label	Operator's Belief
B_1	$WTC(car)$
B_2	$driver(car) = suspect$
B_3	$Near(car, house)$
B_4	$Near(driver(car), house)$
B_5	$Near(house, safehouse)$
B_6	$Has(suspect, item)$
B_7	$Briefcase(item)$
B_8	$Near(suspect, airport)$

Corolla.

$$\mathbf{P}(operator, t_0, WTC(car) \wedge driver(car) = suspect)$$

At some time t_1, the suspect makes a stop at a house.

$$\mathbf{P}(operator, t_1,$$
$$Near(car, house) \wedge Near(driver(car), house))$$

The house also appears to be near the suspected safehouse.

$$\mathbf{P}(operator, t_1, Near(house, safehouse))$$

Some time after the stop, while by the house, the suspect is seen carrying a laptop bag.

$$\mathbf{P}(operator, t_2, Has(suspect, item) \wedge$$
$$Briefcase(item) \wedge Near(item, house))$$

Toward the end of the day, the drone sees the suspect park at a location that is three kilometers from the airport.

$$\mathbf{P}(operator, t_3, Near(suspect, airport))$$

These perceptions are then converted to beliefs as shown in Table 2.

5. Related Work

Much could be said about related work; we shall keep things brief in the present section, and touch upon a few distinctive apsects of our approach, with an eye to our desiderata $\mathcal{D}$.

To start with the general, as most readers well know, nonmontonic/defeasible reasoning and logics that support such reasoning can be traced back a number of decades. McCarthy [33] invented nonmonotonic logics based on circumscription, which is second-order and model-theoretic, the latter aspect at odds with our thoroughgoingly inference-by-inference-schemata approach (at odds with d_4 and d_5). In a firmly argument-centric orientation that specifically commits to the internal structure of arguments as crucial (satisfying d_4), Pollock later invented and implemented the defeasible-logic reasoner OSCAR [34–36] that inspired us.[e] Early seminal work in nonmonotonic logic was carried out (temporally speaking) alongside McCarthy; for instance, specifically, we have the default logics of Reiter [37], in which epistemic possibilities hold in default of information to the contrary. However, none of the excellent work by these three pioneers was intensional in nature; no intensional operators, let alone intensional inference schemata, are to be found (failure of d_6). We leave aside in the interest of economy further assessment of the McCarthy-Reiter-Pollock trio w.r.t. to other desiderata. We do mention that more recently, Licato [38] modeled a complex case of deceptive reasoning and planning from the award-winning television series *Breaking Bad* using default logic. Their work did in fact use a cognitive calculus in the family $\mathscr{C}$ (the Cognitive Event Calculus, $\mathcal{CEC}$, which is devoid of deontic operators) to model the beliefs and intentions of various agents, but didn't have a formalism for assigning strengths to beliefs, and was in the realm of deductive logics, not inductive ones; therefore, while commendable on many fronts, their system does not satisfy d_3.

What about more work in defeasible argumentation systems, considering the promised eye to the desiderata we have laid down? We mention two pieces of impressive prior work, neither of which significantly overlaps our new approach, as we explain:

(1) [39] presents a general framework for structured argumentation, and the framework is certainly computational in nature. The framework, ASPIC$^+$, is in fact Pollockian in nature, at least in part. More specifically this framework is based upon two fundamental principles, the second of which is that "arguments are built with two kinds of inference rules: strict, or deductive rules, whose premises guarantee their conclusion, and defeasible rules, whose premises only create a presumption in favor of their conclusion" [p. 31, [39]]. This second principle is directly at odds with desideratum d_5. In our our intensional inductive calculi, including specifically $\mathcal{IDCEC}$, all non-deductive inference schemata are formally checkable, in exactly the way that deductive inference schemata are. For instance, if some inference is analogical in nature, as long as the schema $\frac{\Phi}{C}$ (Φ for a collection of premises and C for the conclusion) for an analogical inference is correctly followed, the inference is watertight, no different even than even *modus ponens*, where of course specifically we have $\frac{\phi \rightarrow \psi, \phi}{\psi}$.[f]

(2) [41] is an overview of implementations of formal-argumentation systems. However, the overview is highly constrained by two attributes. The first is that their emphasis is on Turing-decidable reasoning problems (at odds thus with d_7). As to the second attribute, the authors are careful to say that their work is constrained by the "basic requirement" that "conflicts" between arguments are "solved by selecting subsets of arguments," where "none of the selected arguments attack each other." Both of these attributes are rejected in our approach. In fact, with respect to the first, most of the interesting parts of automated-reasoning science and technology for us only *start* with problems at the level of the *Entscheidungsproblem*; see in this regard desideratum d_7. As to the second attribute, it too is not true of our approach; in fact, adjudication for us is most needed when there is a complete absence of a state-of-affairs in which no arguments attack each other.

Work in testimonial evidence has a close tie to epistemology, and under that topic a tie

[e] Pollock has unfortunately passed away. Bringsjord's RAIR Lab currently maintains and offers OSCAR, courtesy of resurrection of the (Common Lisp) code by Kevin O'Neill. Today, there are strikingly few extant nattural-deduction automated deductive reasoners, and OSCAR is one at the level of at least $\mathscr{L}_1$.

[f] For a discussion of this sort of explicit rigidity in the case of analogical inference, see [40].

specifically to notions of trust, deontological justification of belief, judgement of character, and evidentialism. In our view, in this regard, helpful prior work is the account of occurrent trust in [42]; the account states that an agent $\mathfrak{a}_b$ trusts an agent $\mathfrak{a}_c$ to do an action α with respect to some goal ψ if and only if:

(1) Agent $\mathfrak{a}_b$ has the goal ψ.
(2) Agent $\mathfrak{a}_b$ believes that:

- Agent $\mathfrak{a}_c$ is capable of doing action α.
- By agent $\mathfrak{a}_c$ doing α, it will ensure ψ.
- Agent $\mathfrak{a}_c$ intends to perform action α.

As to deontological justification of belief, the idea here is that an agent $\mathfrak{a}_h$ is justified in believing ϕ if and only if $\mathfrak{a}_h$ is not obligated to refrain from believing ϕ. Since $\mathcal{IDCEC}^*$ is a deontic cognitive calculus, it can be easily expressed by us as:

$$\frac{S(\mathfrak{a}_s, \mathfrak{a}_h, t, \phi) \qquad \neg O(\mathfrak{a}_h, t, \neg B^n(\mathfrak{a}_h, t, \phi), \chi)}{B^1(\mathfrak{a}_h, t, \phi)} \; I_B^D$$

As to evidentialism, [43] is certainly relevant, and when cast into our formal machinery states that an agent $\mathfrak{a}_h$ is justified in believing ϕ if and only if the belief of ϕ fits the evidence available; more precisely:

$$\frac{S(\mathfrak{a}_s, \mathfrak{a}_h, t, \phi) \qquad \{\phi_1, \ldots, \phi_m\} \vdash \phi \qquad \{\phi_1, \ldots, \phi_m\} \nvdash \bot}{B^1(\mathfrak{a}_h, t, \phi)} \; I_B^E$$

Finally, there is informal but remarkable work on testimonial evidence and character that has inspired us, and which will continue to do so; this is the work of Walton [44,45]. We return to this at the very end of the present paper.

6. Testimony-based Inferencing

Let us denote the agent or operator watching the video streams while gathering and adjudicating beliefs (following notation introduced earlier) as the adjudicator $\mathfrak{a}_{adj}$. Outside intelligence may come from satellite imagery, intercepted radio communications, sources at the site, etc. It is up to the adjudicator to determine whether it believes a *testimony* coming from an outside source.

6.1. *Inference Schemata for Testimonial Evidence*

In $\mathcal{IDCEC}$, testimonies are communicated using the **S**ays operator. (In this cognitive calculus, no NLU or NLG based on logically controlled natural language is used, so this operator is not associated with subsidiary computation for NLP in the present paper.) For example, a testimony from intelligence analyst $\mathfrak{a}_i$ for a claim ω can be represented as $\mathbf{S}(\mathfrak{a}_i, \mathfrak{a}_{adj}, t, \omega)$. That that a formula of this type conforms to the basic triadic structure of testimonies given earlier in the present paper.

We employ a confessedly naïve inference schema for handling testimonies, according to which the operator simply believes everything the intelligence analyst says, at the level of *highly likely* (we end the paper by pointing toward more sophisticated schemata):

$$\frac{\mathbf{S}(\mathfrak{a}_i, \mathit{operator}, t, \omega)}{\mathbf{B}^3(\mathit{operator}, t, \omega)} \; I_{naive}$$

7. Case Study Part II

Given the necessary formal background, we return to the scenario.

7.1. *Intelligence Reports*

Table 3: Beliefs from Testimonies (at likelihood 3 from I_{naive})

Label	Operator's Belief
B_9	$Collab(iperson_1, ISIS\ K)$
B_{10}	$\mathbf{I}(iperson_1, t', bomb(airport))$
B_{11}	$iperson_1 = driver(icar)$
B_{12}	$WTC(icar)$
B_{13}	$Near(iperson_2, safehouse)$
B_{14}	$Collab(iperson_2, ISIS\ K)$
B_{15}	$Has(iperson_2, iitem)$
B_{16}	$Explosive(iitem)$
B_{17}	$Near(driver(icar), safehouse)$
B_{18}	$Near(icar, safehouse)$

There are three pieces of intelligence that the operators received throughout the day. To add granularity, we assigned different intelligence analysts to each of them.

The first piece of intelligence, which comes from prior attacks is that a driver of a white Toyota Corolla is a collaborator of ISIS K and intends to bomb the airport.

$$\begin{array}{ll} Collab(iperson_1, ISIS\ K) & \wedge \\ \mathbf{I}(iperson_1, t', bomb(airport)) & \wedge \\ driver(icar) = iperson_1 & \wedge \\ WTC(icar) & \end{array} \tag{ψ_1}$$

The next piece of outside information comes from intercepted communications and states that a meeting to hand off explosions to an ISIS K collaborator is happening at the safehouse.

$$\begin{array}{ll} Near(iperson_2, safehouse) & \wedge \\ Collab(iperson_2, ISIS\ K) & \wedge \\ Has(iperson_2, iitem) & \wedge \\ Explosive(iitem) & \end{array} \tag{ψ_2}$$

The last piece of outside intelligence comes from a satellite analyst which states that a white Toyota Corolla left the safehouse.

$$\begin{array}{ll} Near(driver(icar), safehouse) & \wedge \\ Near(icar, safehouse) & \end{array} \tag{ψ_3}$$

These pieces of outside intelligence are then converted to beliefs by the operator into Table 3.

Amidst uncertainty, the military need some way to associate objects which are potentially the same. In the next subsections, we will introduce the notion of cognitive transitive nearness as well as the Identity of Indiscernables under Uncertainty.

7.2. *Nearness Properties*

It is clear that the Predicate *Near* has the symmetric property:

$$\forall p_1, p_2 : Near(p_1, p_2) \iff Near(p_2, p_1)$$

The same cannot be said of transitivity. Otherwise, one can chain enough objects that are near each other to not satisfy Nearness. However, it is cognitively plausible that the transitive property of nearness holds, albeit with slightly less confidence with each chain.

$$\begin{aligned}\forall p_1, p_2, p_3 : \ & \mathbf{B}^{\sigma_1}(\mathfrak{a}, t, Near(p_1, p_2)) \land \\ & \mathbf{B}^{\sigma_2}(\mathfrak{a}, t, Near(p_2, p_3)) \implies \\ & \mathbf{B}^{max(0, min(\sigma_1, \sigma_2)-1)}(\mathfrak{a}, t, Near(p_1, p_3))\end{aligned}$$

7.3. *Identity of Indiscernables under Uncertainty*

The standard Identity of Indiscernables state that two objects share all the same properties iff they are the same object as further described in [46] and [8].

$$\forall F : (Fx \iff Fy) \iff x = y$$

Now a version that allows for uncertainty is that it is believable that two objects are the same if you believe they're near each other and share two properties. The following statement is in third-order logic.

$$\begin{aligned}\exists F, G\ \forall x, y : \ & F \neq G \land F \neq Near \land G \neq Near \land \\ & \big(\mathbf{B}^{\sigma_1}(\mathfrak{a}, t, Near(x, y)) \land \\ & \ \mathbf{B}^{\sigma_2}(\mathfrak{a}, t, Fx) \land \mathbf{B}^{\sigma_3}(\mathfrak{a}, t, Fy) \land \\ & \ \mathbf{B}^{\sigma_4}(\mathfrak{a}, t, Gx) \land \mathbf{B}^{\sigma_5}(\mathfrak{a}, t, Gy)\big) \implies \\ & \ \mathbf{B}^{max(0, min(\sigma_i)-1)}(\mathfrak{a}, t, x = y)\end{aligned}$$

7.4. *Associating Objects*

We first want to associate that the two people specified in the outside intelligence refers to the same person. We can do this by first using the symmetric and cognitive transitive nearness properties:

$$\begin{aligned}& \big(\mathbf{B}^3(operator, t, Near(iperson_1, safehouse)) \land \\ & \ \mathbf{B}^3(operator, t, Near(safehouse, iperson_2))\big) \implies \\ & \ \mathbf{B}^2(operator, t, Near(iperson_1, iperson_2))\end{aligned}$$

Then we can use the Identity of Indiscernables under Uncertainty:

$$\begin{aligned}& \big(\mathbf{B}^2(operator, t, Near(iperson_1, iperson_2)) \land \\ & \ \mathbf{B}^3(operator, t, Collab(iperson_1, ISIS\ K) \land \\ & \ \mathbf{B}^3(operator, t, Collab(iperson_2, ISIS\ K) \land \\ & \ \mathbf{B}^3(operator, t, \mathbf{I}(iperson_1, bomb(airport)) \land \\ & \ \mathbf{B}^3(operator, t, \mathbf{I}(iperson_2, bomb(airport)))\big) \implies \\ & \ \mathbf{B}^1(operator, t, iperson_1 = iperson_2)\end{aligned} \tag{B_{19}}$$

Next we want to show a belief that the suspect driving the white Toyota Corolla is driving the same white Toyota Corolla from the pieces of outside intelligence. We first use the nearness property to establish a belief that the suspect's car is near the safehouse.

$$\begin{aligned}&\big(\mathbf{B}^5(operator, t_2, Near(car, house)) \wedge \\ &\mathbf{B}^5(operator, t_2, Near(house, safehouse))\big) \implies \\ &\mathbf{B}^4(operator, t_2, Near(car, safehouse))\end{aligned} \tag{B_{20}}$$

We then use the cognitive transitive property of nearness again and symmetric property to establish that the suspect's car is near the car from the intelligence reports.

$$\begin{aligned}&\big(\mathbf{B}^4(operator, t_2, Near(car, safehouse)) \wedge \\ &\mathbf{B}^3(operator, t_2, Near(safehouse, icar))\big) \implies \\ &\mathbf{B}^2(operator, t_2, Near(car, icar))\end{aligned} \tag{B_{21}}$$

With the two cars near each other, we can then use the fact that they're both white Toyota Corolla's to infer a belief that they're the same object.

$$\begin{aligned}&\big(\mathbf{B}^2(operator, t, Near(car, icar)) \wedge \\ &\mathbf{B}^5(operator, t_2, White(car)) \wedge \\ &\mathbf{B}^3(operator, t_2, White(icar)) \wedge \\ &\mathbf{B}^5(operator, t_2, Corolla(car)) \wedge \\ &\mathbf{B}^3(operator, t_2, Corolla(icar))\big) \implies \\ &\mathbf{B}^1(operator, t_2, car = icar)\end{aligned} \tag{B_{22}}$$

We can then apply the function *driver* to induce:

$$\begin{aligned}&\mathbf{B}^1(operator, t_2, driver(car) = driver(icar)) \implies \\ &\mathbf{B}^1(operator, t_2, suspect = iperson_1)\end{aligned} \tag{B_{23}}$$

It is at this point that the military can infer (via substitution) albeit with a low degree of confidence that the suspect is capable and intends to bomb the airport.

$$\begin{aligned}&\big(\mathbf{B}^5(operator, t_3, Near(suspect, airport)) \wedge \\ &\mathbf{B}^1(operator, t_3, Has(suspect, iitem)) \wedge \\ &\mathbf{B}^1(operator, t_3, Explosive(iitem))\big) \implies \\ &\mathbf{B}^1(operator, t_3, Capable(suspect, t_3, bomb(airport))\end{aligned} \tag{B_{24}}$$

7.5. *Lack of Time Argument*

Since the level of belief in B_{24} is low, a strike will not be permitted unless a belief *beyond reasonable doubt* is held that there is no additional time to counterattack.

ϕ_1 From prior attacks, the operator has a *very likely* belief that the explosive item is either a suicide vest (SVest) or a rocket.

- $B^3(operator, t_3, SVest(iitem) \vee Rocket(iitem))$

ϕ_2 SVests are explosives that can fit into a briefcase.

- $\forall x : SVest(x) \implies Explosive(x) \wedge Briefcase(x)$

ϕ_3 Rockets would not be able to fit into a briefcase.

- $\forall x : Rocket(x) \implies \neg Briefcase(x)$

B_{25} We know from B_7, B_{16} that the suspect has an item that is explosive and an item that fits in a briefcase.

- $\mathbf{B}^3(operator, t_3,$
  $Explosive(iitem) \wedge Briefcase(item))$

B_{26} Through cognitive transitive nearness:

- $\mathbf{B}^2(operator, t_3, Near(item, iitem))$

B_{27} Using the Identity of Indiscernables under Uncertainty we can derive a belief that $iitem = item$ using the relations $Has(suspect, x)$, $ObtainedNearSafehouse(x)$ with a belief level of 2.

- $\mathbf{B}^2(operator, t_3, iitem = item)$

B_{28} The explosive *item* fits inside a briefcase, therefore through disjunction syllogism, it must be an SVest.

- $\mathbf{B}^1(operator, t_3, Svest(item))$

ϕ_4 It is known that SVests are hard to counterattack in a populated area unless the suspect is enclosed.

- $\exists\ suspect, item :$
  $\big(Populated(suspect) \wedge SVest(item) \wedge$
  $Has(suspect, item) \wedge \neg Enclosed(suspect)\big) \implies$
  $\neg Capable(operator, t', counterattack)$

B_{29} The operator perceived the suspect park in a populated and enclosed location. (From perception and $I^s_{\mathbf{P}}$)

- $\mathbf{B}^5(operator, t,$
  $Populated(suspect) \wedge Enclosed(suspect))$

B_{30} The operator held a belief through perception that the suspect would soon be not enclosed.

- $\mathbf{B}^5(operator, t', \neg Enclosed(suspect))$

B_{31} Once the suspect is not enclosed, opportunity would've been lost to counterattack. Therefore, there is no future time available to counterattack.

- $\mathbf{B}^1(operator, t_3, \neg\exists t' > t_3 :$
  $Capable(operator, t', counterattack))$

7.6. *Simulations Achieved via Automated Reasoning*

ShadowProver [47] is an automated reasoner for the (purely deductive) $\mathcal{DCEC}$. In this work, we utilize a nascent automated reasoner — ShadowAdjudicator [48] — which contains a novel algorithm for reasoning about *inductive* cognitive calculi such as $\mathcal{IDCEC}$. It builds directly off of ShadowProver to enable reasoning about formulae with likelihood values.

ShadowAdjudicator is able to automatically find all of the arguments presented herein. The two main arguments, presented in §7.4 and §7.5, are each split into three sub-arguments.

The three sub-arguments of the first argument are: (1) the two people specified in the intelligence reports are the same person, therefore (2) the (perceived) suspect is the same person as the person identified in the intelligence reports, therefore (3) that the suspect is capable of bombing the airport. Those three sub-arguments are found by ShadowAdjudicator in 4.2s, 1.1s, and 0.3s respectively.

The three sub-arguments of §7.5 are: (1) the (perceived) item is the same item as the one mentioned in the intelligence reports, therefore (2) the item is a SVest, therefore (3) there is not enough time to wait to perform a counterattack. These three sub-arguments are found by ShadowAdjudicator in 0.27s, 0.34s, and 0.30s respectively.

8. Concerns; Replies

Before concluding the paper, we devote the present, short section to replies to a pair of concerns we anticipate arising in the minds of some thoughtful readers.

8.1. *What About Prior Logic-based Modeling?*

A reader might reasonably say: "You do have a section above on related work, which is appreciated, but there you focus on a general class of AI systems (into which yours falls), rather than on any specific modeling efforts in logic-based AI. Isn't there some modeling (and simulation) work in the past that is relevant to what you do here? After all, many folks have used formal logic to model various phenomena. For instance, Hayes [49] rather famously long ago modeled the behavior of fluids (in naïve-physics fashion), and much more recently Shanahan [50] modeled the cracking of an egg, which he rightly regards to be a 'benchmark' problem. How does your framework relate to this kind of impressive work?"

We have much respect for the work cited by this imagined reader. In fact we believe that Hayes inaugurated this line of work in nothing short of seminal fashion. However, this work is fundamentally different than our framework, and the work that has bee carried out to erect is. The differences are numerous; we have space here to mention but two; they are as follows. One, our research, our logico-mathematical formalisms (e.g. our cognitive calculi, including the one employed herein), and our automated-reasoning technology; all of this is *intrinsically intensional*. We are not interested the logicist modeling and simulation of objects and processes that are non-cognitive. Now of course it might be said that the cracking of an egg can be quite cognitive, but the fact is that in the aforementioned [50] there is no use of, and more importantly no need for, logics that have a singularly robust lineup of intensional/modal operators.[g] Our second point is this: Our framework is in no way given as first and foremost a contribution to modeling. No, our purpose is to engineer a framework in which AI can make decisions that save lifes, by adjudicating arguments regarding life-and-death questions.

8.2. *Is Your Framework Extensible?*

A second concern can be expressed thus: "Surely you agree that if your framework were a one-off affair, it would have precious little value. So, you must see that a skeptic would demand from you some assurance that your framework will carry over not only to other case studies, but to an entire class of multi-agent decision-making challenges of great consequence."

In reply, we cheerfully admit that our framework is intended to be extensible, and applicable well beyond the particular scenarios we have selected. (We also acknowledge that the work on cracking an egg cited above is very much intended to be applicable to a wide swathe of *physical* phenomena.) But does not one glance at the nature of for instance $\mathcal{IDCEC}$ reveal instantly the broad scope of our framework? After all, the inference schemata at the heart of all of our cognitive calculi are formal, abstract, and entirely domain-general. Indeed, it's very hard for us to fathom how, given the nature of these schemata, anyone rational could fail to see the extensibility and broad applicability of our framework. Consider perhaps the simplest inference schema known: *modus ponens*, which made a brief appearance above:

$$\frac{\phi \rightarrow \psi, \phi}{\psi}$$

[g]It is probably worth pointing out that in our cognitive calculi said to be "event calculi," the event calculus (which was invented by Shanahan himself; see e.g. [51]) is not used in its original form, but is rather cast in cognitive form. This has been the case going back to the very first cognitive calculus invented and used in modeling-and-simulation work: [15].

How can it fail to be that this inference schema is applicable to any number of spheres, and that it can be part of larger and larger collections of inference schemata? Likewise, the intensional inference schemata in $\mathcal{IDCEC}$ have the same wide scope, and can be parts of arbitrarily extended collections of inference schemata.

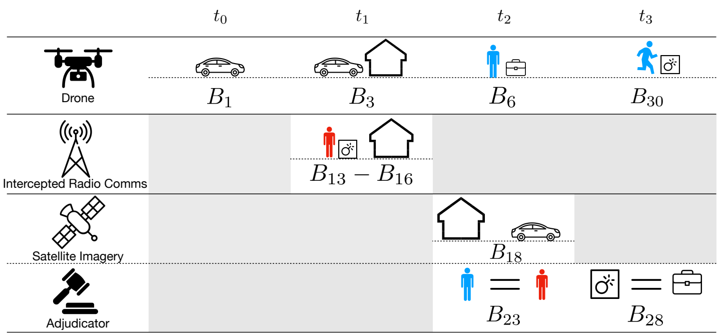


Fig. 1. Scenario

9. Conclusion & Next Steps

The scenario ends with a low degree of confidence that there was no additional opportunity to counterattack. This means that the two possible conditions to permit a strike were both unsatisfied: viz.,

- a belief *beyond reasonable doubt* that the suspect was capable and intending to bomb the airport; and
- a positive belief that the suspect was capable of and intending to bomb the airport, and a belief *beyond reasonable doubt* that there is no additional time to counterattack.

These conditions address the principle of distinction in the DoD law of war manual while also addressing the fog of war or uncertainty frequently present in warzones.

A cognitively plausible argument presenting a positive belief that the suspect was capable of and intent on bombing the airport, given the foregoing, is seen to be justified. This argument made use of inductive inference, which in turn more generally made use of higher-order logic in order to associate objects via both our "uncertainty-ized" Identity of Indiscernables, and transitivity of nearness.

Since an argument couldn't be made at the *beyond reasonable doubt* level that there was no additional time to counterattack, from a purely rational viewpoint we respectfully assert that more time should've been taken to track the suspect into the future, and gather newly arriving information. We believe that as logicist AI of the type we have featured herein progresses, the automated adjudication of arguments can extend beyond the technology that we exhibit herein, and take into account the blast radius of the missile, time between permitting a strike and it landing, velocity of the target, and so on, in order to maximize observation time and the accuracy of automated human/AI reasoning about agents of interest, and potential dangers arising therefrom.

To sum up, the main purpose of our work has been to provide to artificial agents an automated representation-and-reasoning capability sufficiently expressive and powerful to deal with ethically charged cases like the Kabul tragedy (which makes its expressive and automated-reasoning reach unprecedented, to our knowledge) — and this purpose has been achieved. Our chief overarching technical result is invention of the inductive logic we have displayed, its use for a robust and relevant case study, and its automaticity and automated runs. Our AI automatically finds in this case study the relevant proofs and — since this is inductive logic, not deductive — arguments, , and we are the first group to achieve any such engineering, as far as we know.

What about next steps along the line of investigation described above? As alert and discriminating readers will doubtless have detected, the brute fact is that our inference schema for testimonial evidence is naïve. The next phase of our efforts will be to complete the specification of a robust and credible inference schema in a variant of $\mathcal{IDCEC}$, and achieve implementation via our automated reasoners.

Fortunately, there is some seminal prior work on such evidence under the umbrella of *in*formal logic, carried out by Walton [44,45]. Taking, we admit, considerable liberties in sharpening this work so as to make use of it in our formalisms and in the automated reasoners that enable such agents to compute, we can lay out at least a provisional formal version of one such inference schema that can be expressed in the inductive cognitive calculus $\mathcal{IDCEC}$ (see Figure 2), and with a presentation of it immediately below we end.

$$\frac{\begin{array}{c}\Vdash \mathbf{B}^{\sigma_1}(\mathfrak{a}_{adj}, EpistemicPos(\mathfrak{a}_w, \phi)) \\ \Vdash \mathbf{B}^{\sigma_2}(\mathfrak{a}_{adj}, \geq (character(\mathfrak{a}_w), k)) \\ \mathbf{K}(\mathfrak{a}_{adj}, \mathbf{S}(\mathfrak{a}_w, \phi, \mathfrak{a}_{adj}))\end{array}}{\mathbf{B}^{f(\sigma_1,\sigma_2,k)}(\mathfrak{a}_{adj}, \phi)}$$

Fig. 2. A Provisional Inference Schema for Testimonial Evidence. *The traditional single-turnstyle $\vdash$ for straight deductive provability is here replaced with a variant that indicates that what follows it is the conclusion of inference that may be either deductive or inductive, expressed, respectively, by a proof or formally valid argument.*

Acknowledgments

The work reported herein exploits key automated-reasoning theory & technology developed via multi-year support from both AFOSR (Award # FA9550-17-1-0191) and ONR of investigators Bringsjord & Govindarajulu, and of doctoral researchers (including Giancola) under their supervision in the Rensselaer AI *&* Reasoning (RAIR) Laboratory. B&G are especially grateful for ONR support (Award # (N000141912558) intended to surmount Arrow's Impossibility Theorem (AIT) via the brand of artificial adjudication described above — and this pair hereby asserts that such surmounting (courtesy in no small part of their requiring that "voting" must be supported by an argument) has come to pass. They are also grateful for ONR support under Award # N00014-17-1-2115 for the development of logic-based machine learning, which in part makes the formal schemes and reasoning therein shown herein possible.

References

1. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (Pearson, New York, NY, 2020). Fourth edition.
2. S. Bringsjord, The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field Unto Itself, *Journal of Applied Logic* **6**, 502 (2008) `http://kryten.mm.rpi.edu/SB\_LAI\_Manifesto\_091808.pdf`.
3. H. Levesque and G. Lakemeyer, Chapter 24: Cognitive Robotics, in *Handbook of Knowledge Representation*, (Elsevier, Amsterdam, The Netherlands, 2007).
4. M. Ashcraft and G. Radvansky, *Cognition* (Pearson, London, UK, 2013). This is the 6th edition.
5. S. Z. Hashemi, R. Faiez, L. C. Baldor and J. Krauss, Kabul airport attack kills 60 afghans, 13 us troops, *AP News* (2021) `https://apnews.com/article/europe-france-evacuations-kabul-9e457201e5bbe75a4eb1901fedeee7a1`.

6. M. Fitting, Intensional Logic, in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta, 2015) `https://plato.stanford.edu/entries/logic-intensional`.
7. P. Dung, On the Acceptability of Arguments and its Fundamental Fole in Nonmonotonic Reasoning, Logic Programming and N-Person Games, *Artificial Intelligence* **77**, 321 (1995) .
8. H. D. Ebbinghaus, J. Flum and W. Thomas, *Mathematical Logic (second edition)* (Springer-Verlag, New York, NY, 1994).
9. J. Paris and A. Vencovská, *Pure Inductive Logic* (Cambridge University Press, Cambridge, UK, 2015).
10. G. Johnson, *Argument & Inference: An Introduction to Inductive Logic* (MIT Press, Cambridge, MA, 2016).
11. M. Ashcraft, *Human Memory and Cognition* (HarperCollins, New York, NY, 1994).
12. E. B. Goldstein, *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience* (Cengage Learning, Boston, MA, 2008). This is the 5th edition.
13. S. Bringsjord, N. S. Govindarajulu, J. Licato and M. Giancola, Learning *Ex Nihilo*, in *GCAI 2020. 6th Global Conference on Artificial Intelligence*, , EPiC Series in Computing Vol. 72 (EasyChair Ltd, Manchester, UK, 2020).
14. 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).
15. K. Arkoudas and S. Bringsjord, Propositional Attitudes and Causation, *International Journal of Software and Informatics* **3**, 47 (2009) `http://kryten.mm.rpi.edu/PRICAI\_w\_sequentcalc\_041709.pdf`.
16. S. Bringsjord and N. Sundar Govindarajulu, Rectifying the Mischaracterization of Logic by Mental Model Theorists, *Cognitive Science* **44**, p. e12898 (2020) `https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12898`.
17. S. Bringsjord, M. Giancola and N. S. Govindarajulu, Logic-Based Modeling of Cognition, in *The Handbook of Computational Psychology*, ed. R. Sun (Cambridge University Press, Cambridge, UK, (forthcoming)) `http://kryten.mm.rpi.edu/Logic-basedComputationalModelingOfCognition.pdf`.
18. S. Bringsjord, Declarative/Logic-Based Cognitive Modeling, in *The Handbook of Computational Psychology*, ed. R. Sun (Cambridge University Press, Cambridge, UK, 2008) pp. 127–169. This URL goes to a preprint only. `http://kryten.mm.rpi.edu/sb\_lccm\_ab-toc\_031607.pdf`.
19. K. Arkoudas and S. Bringsjord, Vivid: An AI Framework for Heterogeneous Problem Solving, *Artificial Intelligence* **173**, 1367 (2009) `http://kryten.mm.rpi.edu/KA\_SB\_Vivid\_offprint\_AIJ.pdf`.
20. S. Bringsjord, N. S. Govindarajulu and M. Giancola, Automated Argument Adjudication to Solve Ethical Problems in Multi-Agent Environments, *Paladyn, Journal of Behavioral Robotics* **12**, 310 (2021) `https://doi.org/10.1515/pjbr-2021-0009`.
21. G. Gentzen, Investigations into Logical Deduction, in *The Collected Papers of Gerhard Gentzen*, ed. M. E. Szabo (North-Holland, Amsterdam, The Netherlands, 1935) pp. 68–131. This is an English version of the well-known 1935 German version. .
22. D. Prawitz, The Philosophical Position of Proof Theory, in *Contemporary Philosophy in Scandinavia*, eds. R. E. Olson and A. M. Paul (Johns Hopkins Press, Baltimore, MD, 1972) pp. 123–134 .
23. P. Schroeder-Heister, Proof-Theoretic Semantics, in *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta, 2018) `https://plato.stanford.edu/entries/proof-theoretic-semantics`.
24. S. Bringsjord, J. Hendler, N. Govindarajulu, R. Ghosh and M. Giancola, The (Uncomputable!) Meaning of Ethically Charged Natural Language, for Robots, and Us, From Hypergraphical Inferential Semantics, in *Trustworthy Artifical-Intelligent Systems*, ed. I. Ferreira (Springer, Cham, Switzerland, (forthcoming)) This URL goes to a preprint only. `http://kryten.mm.rpi.edu/UncomputableNLURobots032421.pdf`.
25. N. Francez, *Proof-theoretic Semantics* (College Publications, London, UK, 2015).
26. T. Anderson, D. Schum and W. Twining, *Analysis of Evidence* (Cambridge University Press, Cambridge, UK, 2009). This is the 3rd edition.
27. J. Wigmore, *Science of Judicial Proof, as Given by Logic, Psychology, and General Experience and Illustrated in Judicial Trials* (Little, Brown and Co., Boston, MA, 1937).
28. S. E. Preston and R. S. Taylor, *Department of Defense Law of War Manual*, tech. rep., General Counsel of the Department of Defense Washington United States (2016).
29. N. Govindarajulu, S. Bringsjord, R. Ghosh and M. Peveler, Beyond the Doctrine of Double

Effect: A Formal Model of True Self-Sacrifice, in *Robots and Well-Being*, eds. M. Ferreira, J. Sequeira, G. Virk, M. Tokhi and E. KadarIntelligent Systems, Control and Automation: Science and Engineering (Springer, Basel, Switzerland, 2019) pp. 39–54. The URL here is to a rough preprint. `http://kryten.mm.rpi.edu/NSG\_SB\_RG\_MP\_DDE\_SelfSac\_110617.pdf`.
30. S. Bringsjord, N. Govindarajulu and M. Giancola, Automated Argument Adjudication to Solve Ethical Problems in Multi-Agent Environments, *Paladyn, Journal of Behavioral Robotics* **12**, 310 (2021), The URL here goes to a *rough*, *uncorrected*, *truncated* preprint as of 071421. `http://kryten.mm.rpi.edu/AutomatedArgumentAdjudicationPaladyn071421.pdf`.
31. M. Aikins, C. Koettl, E. Hill and E. Schmitt, NY Times Investigation: "In U.S. Drone Strike, Evidence Suggests No ISIS Bomb", *New York Times* (2021) `https://www.nytimes.com/2021/09/10/world/asia/us-air-strike-drone-kabul-afghanistan-isis.html`.
32. S. D. Said and J. F. Kirby, Pentagon Press Secretary John F. Kirby and Air Force Lt. Gen. Sami D. Said Hold a Press Briefing(November 2021).
33. J. McCarthy, Circumscription—A Form of Non-Monotonic Reasoning, *Artificial Intelligence* **13**, 27 (1980) .
34. J. Pollock, Defasible Reasoning with Variable Degrees of Justification, *Artificial Intelligence* **133**, 233 (2001) .
35. J. L. Pollock, How to Reason Defeasibly, *Artificial Intelligence* **57**, 1 (1992) `citeseer.ist.psu.edu/pollock92how.html`.
36. J. Pollock, *Cognitive Carpentry: A Blueprint for How to Build a Person* (MIT Press, Cambridge, MA, 1995).
37. R. Reiter, A Logic for Default Reasoning, *Artificial Intelligence* **13**, 81 (1980) .
38. J. Licato, Formalizing Deceptive Reasoning in Breaking Bad: Default Reasoning in a Doxastic Logic, in *2015 AAAI Fall Symposium Series*, 2015.
39. S. Modgil and H. Prakken, The ASPIC$^+$ Framework for Structured Argumentation: A Tutorial, *Argument & Computation* **5**, 31 (2014) .
40. S. Bringsjord and J. Licato, By *Dis*analogy, Cyberwarfare is Utterly New, *Philosophy and Technology* **28**, 339 (2015) `http://kryten.mm.rpi.edu/SB\_JL\_cyberwarfare\_disanalogy\_DRIVER\_final.pdf`.
41. F. Cerutti, S. A. Gaggl, M. Thimm and J. Wallner, Foundations of Implementations for Formal Argumentation, in *The IfCoLog Journal of Logics and their Applications; Special Issue Formal Argumentation*, eds. P. Baroni, D. Gabbay, M. Giacomin and L. Van der Torre, (8) (College Publications, 2017) pp. 2623–2705 .
42. A. Herzig, E. Lorini, J. F. Hübner and L. Vercouter, A logic of trust and reputation, *Logic Journal of the IGPL* **18**, 214 (2010) .
43. R. Feldman and E. Conee, Evidentialism, *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* **48**, 15 (1985) .
44. D. Walton, *Witness Testimony Evidence: Argumentation, Artificial Intelligence, and Law* (Cambridge University Press, Cambridge, UK, 2008).
45. D. Walton, *Character Evidence: An Abductive Theory* (Springer, Dordrecht, The Netherlands, 2010).
46. P. Forrest, The Identity of Indiscernibles, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2020) Winter 2020 edn. .
47. 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`.
48. 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*.
49. P. Hayes, The Naïve Physics Manifesto, in *Expert Systems in the Microeletronics Age*, ed. D. Mitchie (Edinburgh University Press, Edinburgh, Scotland, 1978) pp. 242–270 .
50. M. Shanahan, An Attempt to Formalise a Non-Trivial Benchmark Problem in Common Sense

Reasoning, *Artificial Intelligence* **153**, 141 (2004) .
51. M. Shanahan, The Event Calculus Explained, in *Artificial Intelligence Today (LNAI 1600)*, eds. M. Wooldridge and M. Veloso (Springer, New York, NY, 1999) pp. 409–430 .

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.015

ETHICAL DESIGN OF EMBODIED AUTONOMOUS SYSTEMS

M. ANTONIETTA GRASSO and DANILO GALLO

Naver Labs Europe - 6, chemin de Maupertuis
Meylan 38240, France
E-mail: antonietta.grasso@naverlabs.com, danilo.gallo@naverlabs.com
www.naverlabseurope.com

With the arrival of intelligent and learning technologies, embodied or not, new questions have arisen around the best fit of the human activities with them. The study of how to best fit the human with the machine to gain productivity has started as early as the first industrialisation steps. With the advent of networks, computers have gone beyond productivity tools and become support for collaborative activities. With their large-scale learning from data, embodiment and sensing extensions, they are now entities that can be defined as *intelligent*, given their capability of mimicking some and exceeding some other human capability. They are proactive in the sense that they can autonomously take the initiative and they are learning in the sense that their behaviour is continuously tuned by their interaction with the environment. These major changes have induced also changes in the way to study the fitting of the human with the machines. In the beginning, it was a matter of controlled lab experiments where features like human perception, movements and memory were studied to understand how to best shape the interaction to gain efficiency. Subsequently, the object of study has become a complex socio-technical entanglement. Facing the complexity of the new object of study new methods have been adopted capable of grasping a picture much closer to the reality of the new type of interaction. In parallel has come the recognition that multiple actors are affected by the introduction of the machines and tensions are typically present about their interests in using them. This methodological extension has been due to the impossibility of isolating specific variables with the method typic of hard sciences inquiries that had been adopted in the laboratory studies. Lately, the HCI community has undergone another extension: while the first studies were oriented toward productivity gain, new objectives started to appear oriented toward supporting *human flourishing* [1]. The objective of usability of computing is clearly not in contradiction with this new objective, but it is rather an element of a larger picture. This larger picture includes a reflection on what is the impact of putting in place a socio-technical autonomous system. When considering a socio-technical perspective the impact needs to account for a multi-stakeholder perspective and short- and long-term aspects [2]. Indeed, the shift to learning systems introduces new concerns already at the usability level, given that their behaviour could be difficult to understand and the principle of consistency could be easily broken. When moving to embodied intelligent systems a physical component is presents which adds an additional dimension to consider for what could evolve, go wrong and leveraged. Behaviour could be more intrusive, but on the other side physical elements could be leveraged as an additional channel to make clear a behaviour or lack of. A system that is exhibiting a complex and evolving behaviour (i.e. which is *AI-infused*, according to [3]) therefore needs to make use of a number of elements: 1) make clear to the user what are its capabilities, 2) use the sensing to understand the specific situation and adapt to it, 3) recognise when wrong and put in place repair measures, 4) demonstrate how it is evolving [3]. These guidelines have been defined upon examination of more than 20 years of studies and the direct examination of 13 products [3]. This is the minimal set of requirements that usability demands for this new generation of technology, embodied or not. However, as mentioned earlier, the more powerful and adaptive behaviour this technology exhibits, the more an examination of its impact is demanded, to avoid harm and possibly produce individual and societal good. Technology has never been as much under scrutiny as recently, with a number of guidelines and assessment methods being disseminated. Ethics is about humankind identifying what values are fundamental and worth defending. This can be done in a protective way to avoid harm, or in a propositive way to promote the good. Additionally, values that are contextual, i.e. more specific to a certain group of stakeholders, are also present. An example of the first type is the EU guidelines HLEG [4], an example of the second type are the UN guidelines for sustainable growth [5], while contextual values can be derived case by case with methods like Value Sensitive Design [2] and an example is given in [8].

While guidelines about the proposed values have been defined, the work about mapping those to technical solutions remains still scattered and highly fragmented in various subdomains. We have done a preliminary review of the body of work and identified three subcategories of area of intervention. The first pertains to the collection and use of the data. The new statistically based models require large datasets while sensing, and user implicit and explicit feedback can fine-tune and enrich the system functionality. However, how these data are collected and used requires user awareness. Additionally, in the case of foundational models, these data can originate new data that can be biased and present toxic and harmful elements. The second major area of intervention relates to the autonomy that is left to the human in face of the intelligent capabilities of the systems. Various aspects of ethical concern that need to be addressed exist. In the case of manipulative systems using *dark patterns* [7], the question is how to leave control to the user when e.g. connecting to social media. At work, there could be issues of deskilling caused by the replacement of human capabilities e.g. when utilising diagnosis systems and also issues of human dignity when technology is enacting algorithmic management over the employees and deciding the order, length and tools they have to use to perform their tasks. Finally, a third category relates to how to support the exhibition of behaviour that can be judged as trustworthy. Trust can be defined as exhibiting consistently over time a competent and not harmful behaviour. While usability issues certainly relate to this topic, it pertains to ethics the design of technology that is not causing harm and prevents confidence loss even if aimed at positive values. Also, it is worth noticing that these areas of intervention are correlated and any analysis of them in a specific system should consider them holistically to identify possible tensions. While in the case of services, design guidelines have been put forward, for autonomous embodied systems the effort is still spread across specific projects and application domains. We believe that a systematic effort is required to start to frame existing design solutions and propose associated patterns and design methodologies. This is an effort we are currently undertaking.

References

1. Paul Dourish. 2019. User experience as legitimacy trap. *Interactions* 26, 6 (November - December 2019), 46–49.
2. Batya Friedman and Peter H. Kahn. 2002. Human values, ethics, and design. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. L. Erlbaum Associates Inc., USA, 1177–1201.
3. Saleema Amershi et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19*). Association for Computing Machinery, New York, NY, USA, Paper 3, 1–13.
4. HLEGAI. (8 April 2019). High-Level Expert Group on Artificial Intelligence, EU - *Ethics guidelines for trustworthy AI.* https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
5. United Nations. Transforming Our World: *The 2030 Agenda for Sustainable Development*. Pub. L. No. A/RES/70/1 (2018). United Nations.
6. Luciano Floridi et al. How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics* 26, 1771–1796 (2020).
7. Arunesh Mathur et al. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* Association for Computing Machinery, New York, NY, USA, Paper 360, 1–18.
8. Tommaso Colombino et al. 2021. Ethical Design of a Robot Platform for Disabled Employees: Some Practical Methodological Considerations. *Frontiers in Robotics and AI* 8:643160.

SECTION–4

SOCIAL ROBOTICS AND SOCIETY

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.003

LEARNING METHOD OF DATA BIAS EMPLOYING MACHINE LEARNING FOR KIDS

HYO-EUN KIM

Hanbat National University, South Korea
E-mail: qualia9@gmail.com

The goal of this paper is to present a specific case of education that recognizes the process of artificial intelligence bias using 'Machine Learning For kids' (https://machinelearningforkids.co.uk/), one of the machine learning platforms. In this paper, by giving examples of specific AI implementation systems, we present the process of constructing the bias that the AI may have on the machine learning platform and the method of experiencing the bias. In general, the method of learning AI ethics is an education and learning method that introduces cases of ethical issues that arise in relation to the development of artificial intelligence and seeks solutions through discussion. These methods are the most basic and universally used educational methods and are learning and educational methods necessary for a basic and general understanding of the contents. The use of the machine learning platform proposed in this paper is a method that makes theoretical contents indirectly practical in addition to basic information delivery education and can be selected and operated independently according to the goals and objectives of education. In particular, in a situation where artificial intelligence is widely used in many fields like today, it can help students who will go into society to strengthen their work skills to be used as office workers. Knowing the ethical mechanism allows you to know exactly what aspects and points are problematic, which areas need to be improved, and which issues need to be addressed.

ICRES 2022: 7th International Conference on Robot Ethics
and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.006

THE CYBORG'S SELF-IDENTITY PROBLEM IN A HYPER-CONNECTED SOCIETY AND ITS ALTERNATIVES[1]

YEONHEE LEE

Institute of Education, Gyonsang National University
Jinju City, Republic of Korea
E-mail: proy13@naver.com
https://www.gnu.ac.kr/

Cyborgs, which mean a combination of humans and robot s or robotized humans, are no longer virtual beings that appear only in science fiction. With the advancement of science and technology, it is expected that various forms of cyborg will already exist and appear anew. On the other hand, modern society is gradually becoming a hyper-connected society in which information and intelligence of everything, such as people and objects, are closely connected in a single organic network and communicate smoothly. In this society, cyborg, a hybrid existence of humans and robot s in this society, can face difficulties in establishing its self-identity in particular. In this context, this study focuses on the problem of cyborg's self-identity that may arise in a hyper-connected society and suggests an alternative.

1. Introduction

The purpose of this study is to overhaul our mental foundation on the path of the new era by examining and reflecting on the human and ethical problems caused by the development of advanced science and technology. With the development of AI (artificial intelligence) and information and communication technology, a society in which information [data] and intelligence of everything, including people and things, are closely connected to one organic network is becoming visible. This society is called the so-called Hyper-connected Society. Many predict that a hype-rconnected society will facilitate choice and decision-making by providing personalized information and generalized intelligence to all members. And it is expected that this will improve the convenience and efficiency of individual lives, and the gap in life (inequality) between individuals will be smaller than now. Besides that, I predict that in such a data-driven society, human actors will inevitably become cyborgs. Here, the cyborg I refer to means a combination of humans and robots or a robotized human. In other words, in a hyper-connected society, anyone will live 'in combination with automated machines'. Cyborgs are no longer virtual beings that only appear in science fiction. This is because with the development of science and technology, various types of cyborgs already exist and are expected to emerge anew.

However, a hyper-connected society that causes human cyborgization does not promise only a bright future. In this context, I would like to discover ethical problems predicted in the future society and consider them ethically. In addition, I would like to propose an ethical alternative to solve this problem.

[1] This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5B5A16077213)

2. Meaning of cyborg

What is a cyborg? Cyborgs often refer to humans combined with machines. For this reason, this existence feels like a virtual existence that appears only in movies and novels. In fact, cyborg is a compound word of cybernetic organism, which first appeared in 1960. American computer expert Mandfred Clynes (1925~2020) and psychiatrist Nathan Kline (1916~1982) first used the word in a paper "Cyborgs and Space" co-published in September that year. In this paper, they named the modified organism Cyborg by applying technology. Since organisms usually refer to creatures with a certain self-regulation ability, the cyborgs they named are not limited to humans. Even when technology is applied to the body of animals such as dogs and cats, it can be called cyborg. [1]

When the meaning of cyborg is defined as a 'technically modified organism', it can be said that the types are very diverse. This is because there may be differences in degree, but there are a wide variety of technologies that can transform the human body and mind. For instance, cyborgs include people with enhanced immunity after being vaccinated against the coronavirus, people who walk with prosthetics on their amputated legs, people who take medicine to strengthen concentration, and people who live with artificial hearts. In addition, people who live by recognizing chips connected to computers in the brain can also be called cyborgs. Any person who has enhanced mental or physical abilities by applying various genetic, pharmaceutical, or mechanical techniques to the biological body can be called a cyborg.

However, considering the fact that our activities today are carried out with the help of a lot of science and technology, the concept of cyborgs can be further expanded. For example, we can rely on glasses to see things more clearly, travel long distances faster with the help of cars, and no longer have to memorize vast amounts of information thanks to smartphones. In this respect, cyborgs are contemporaries we are already meeting, and we are also ourselves living in the age of science and technology.

Expanding the meaning of cyborgs in this way, I would like to point out that in the future, we humans are likely to become cyborgs, especially combined with 'automated machines'. The recent rapid development of computer and robotics and the success of several high-tech technologies are due to the visualization of the emergence of human actors combined with these automated machines. Automated machines generally mean what we call robots. Fundamentally, it is not easy to define a robot that appears by human literary imagination and is being implemented in reality in a single word. This is because they are objects that are being developed even at this moment. For this reason, even robot experts say it is difficult to define the identity in one word.[2] However, if you search for the word robot in *The American Heritage Dictionary of the English Language*, you can see that its dictionary meaning is "1. A mechanical device that some times memories a human and is capable of performing a variety of complex human tasks on the system."[3] In short, robots simply imply 'automated machines that can move without human intervention'. Then, specifically, how will humans be combined with robots in the future?

3. Current development status and Prospects of Cyborg Technology

I said earlier that there is a high possibility that we will become cyborgs combined with automated machines or robots in the future. Let me now state the reason. Today's scientists and policymakers want to create an era in which individuals connect to a central computer network and receive intelligence from information [data] and AI from time to time, while sharing information [data] and intelligence with other connected individuals. They expect that if the information and intelligence of everything, such as people and things, are shared and utilized, individuals' intellectual ability will be greatly improved, and inequality between individuals due to differences in in information and intelligence levels will disappear. A prime example of this expectation is the Smart City construction project promoted by today's scientists and policymakers from each country. Because today, the term 'smart' usually means 'connected by communication, and functioning autonomously based on data collected by sensors', and 'Smart City' means a city that applies information and communication technology to the design, planning, construction and operation of urban infrastructure to ensure that all citizens can solve problems in their daily lives, such as transportation, environment, housing, and facilities, and enjoy a pleasant and convenient life.

As such, a society in which information and intelligence of everything, such as people and things, are closely connected and communicated smoothly on a single organic net is often described as a 'Hyper-connected Society' in other expressions. A hyper-connected society is being visualized by a series of recent advances in science and technology. For example, 5G network technology and Internet of Things connect each individual closely to the surrounding objects like a spider's web, and AI, big data, and sensor technology quickly collects and analyzes a large amount of information about individuals. For this reason, each individual not only intentionally provides his or her personal information by using computers, smartphones, and SNS, but also unknowingly exposes it in his or her daily life. Thus, AI algorithms that somehow grasp an individual ones' secret state, taste, and preference will increasingly know one better than oneself. And this will make it easier for individuals to make choices and decisions within the algorithm's recommended alternatives.

In this respect, I argue that in the future, we will live as cyborgs, a creature combined with robots. Often, the question, "What is a robot?" is reminiscent of machines and devices with physical driving body. However, considering the definition of a robot, the main feature of the robot can be said to be 'automation' rather than a physical body. for this reason, although it is an intangible program, AI can also fall into a kind of robot category. AI literally means mechanical intelligence 'artificially-made' by humans. In the 1950s, the scientific community began in earnest to create AI that imitated the human mind based on computational-functionalism. Computational-functionalism understands the action of the mind as the operation of formal computational systems embodied in the brain. In other words, it is the position that there is no particular difference between the brain's action and the computer's action. In this background, the scientific community has tried to develop AI as a computer algorithm with high computational capabilities. The scientific community has recently achieved remarkable success. That is, computers that remained in the ability to calculate within human-designed programs have been developed to a level where data extracted from things can be recognized, learned, and analyzed alone without human intervention. AI refers to such an automated computer algorithm program that can repeatedly process information on its own without human intervention. Therefore, AI can be classified as a robot even if it does not have any type of body. This is why a software that repeatedly performs a specific task is called a 'bot' similar to a robot.

AI algorithms are expected to influence individual choices and decisions in a variety of areas, from relatively light and simple decisions such as choices in shopping and hobby activities to heavy and significant decisions that determine the direction of individual life, such as political decisions and life-sustaining decisions.[4] Maybe we're already cyborgs. Because, in fact, we are already affected by AI in relation to relatively light decisions. For example, look closely at the advertisement that pops up in the Internet window. We can realize that most of the things are things we need, or things we've always been interested in. Such advertisements are customized advertisements selected and provided by AI algorithms that analyzed my access to the Internet. Thanks to this, we can easily get what we want (knowledge, information, etc.). Life with AI is faster and more efficient than ever. As time goes by, it will be difficult for us to imagine daily life without AI. Cyborgization combined with AI, so to speak, is our naturally anticipated future.

4. Cyborg's Self-Identity Problem

4.1 an anticipated change

Indeed, as many hope and expect, hyper-connected societies may improve the convenience and economic efficiency of individuals and reduce the gap between individuals by providing personalized information and generalized intelligence to make choices and decisions easier. However, this society does not only promise a rosy future. This is because it can be predicted that a hyper-connected society centered on data and intelligence will bring about certain changes in relation to the individual's 'self'. In this chapter, I'd like to discuss the problem of self-identity that cyborgs in the future can often face with these changes. So, before dealing with that, let me first state two changes that many people expect in relation to the self.

4.1.1 The birth of a variable, relative, and pluralistic self

In a hyper-connected society, the self can have variable, relative, and multiple attributes. This is because in a data-intelligence-driven society, the identity of the self is not fixed as one, and can be constantly reconstructed according to the flow of data. In other words, the identity can be formed changeably depending on how **I** combine and separate the information that represents ***myself***. In addition, just as it is possible for a self to create multiple IDs and act as multiple selves in cyberspace, a number of identities with different personalities may be established at the same time.[5]

4.1.2 Boundary collapse between self and other

In a hyper-connected society, as Kurzweil (2000) predicted [6], the boundary between self and others may collapse. In other words, in such a society, individuals' thoughts, emotions, and intellectual abilities connected to one network are integrated and can coexist in an overlapping manner without being distinguished from each other. So, it may be difficult to distinguish between 'my thoughts, feelings, and intelligence' and 'other's thoughts, feelings, and intelligence'. In such a society, even the distinction between what AI has created (fake) and what I have created (real) will become difficult. Accordingly, the self with a unique inner self may no longer be a meaningful concept.

4.2 ethical issues

As expected changes suggest, the future hyper-connected society has the possibility to create new self-concepts that conflict with our familiar concepts of 'self as a fixed and unified entity' and 'self with a unique inner self' or disappear 'individual self'. However, saying that the concept of self as an individual entity is eliminated is the same as saying that the concept of 'individuality' (one's own nature distinguished from other people or individuals) cannot be established. If the concept of individuality becomes invalid in the future, it may be difficult for cyborgized individuals to establish their own unique self-identity. Meanwhile, failure to properly establish their own identity can lead to several serious ethical problems. I named the particularly anticipated ethical issues as follows: (1) *The problem of responsibility, (2) the problem of dignity, and (3) problem of existence*. Let's look at each one.

4.2.1 The Problem of Responsibility

In a hyper-connected society, the problem of responsibility that can be attributed to an individual (individual self) as a moral agent is triggered. In other words, in a hyper-connected society where it is impossible to distinguish one unique self, it is impossible to ask whose idea it is and whose action it is. If so, who is responsible when a moral error occurs?

4.2.2 The problem of dignity

The problem of dignity of members of a hyper-connected society also arises. Since modern times, the concept of individuality has been the basis for human intrinsic value (dignity). However, a society where everything is connected can unify the thoughts and lifestyles of its members. How can human values be guaranteed in such a society that threatens individuality? In other words, on what basis will members of a hyper-connected society be respected for their values?

4.2.3 The problem of existence

The existential problem of members of a hyper-connected society also arises. If one person cannot clearly answer "Who am I (individual identity)," eventually, "How should I live (moral identity)" will also not be clearly known. Then, how should members live in a hyper-connected society? In other words, how can they find the meaning and purpose of life and live a full life without wandering?

5. How should we prepare for the Cyborg Era?

5.1 Suggestion (1) Recognizing the Value of Individuality

The existing view of self that we are familiar with is the so-called 'modern view of self'. This view has dominated our understanding of ourselves since it was founded by Descartes (17th century) and elaborated by Kant (18th century). In the modern view of self, 'self', which means 'I', is generally identified with a subject with a unique subjective consciousness or freewill within an individual. This self is described as having a single, fixed and unified identity centered on one body. At the same time, this perspective highly evaluates the rational ability of the individual self. In other words, in modern culture, reason has been mainly explained as 'the ability to produce decisions without procedural contradictions by calculating or inferring information about the world perceived through sensation according to certain laws'. However, this portrayal of a priori, invariable, independent self and reason with procedural rationality and autonomy raised the value of individuals in the world, while also causing to occur discrimination and inequality against members of society who do not conform to such view of self. [2]

However, as the transition to a hyper-connected society became visible, a new view of self has emerged with the prospect that the existing view of self could no longer be established. Its content (e.g., the negation of the metaphysical self, the prediction of the emergence of a variable, relative, and pluralized self) can be referred to as a post-modern view of self in that it not only has no continuity with the modern view of self, but also conflicts. The post-modern

[2] For example, R. Brideotti(1954~), a representative posthumanist, points out the problem of 'modern human'. 'The modern human' she refers to means a 'rational agent' that modern culture defines as an ideal being. She says humans, normal people, Europeans, men, and white people belong here. On the other hand, she argues that modern culture has always discriminated against and excluded people who lack the attributes of 'reason' or who do not conform to the image of rational agents (non-human, disabled, non-European, women, and people of color).[7]

view of self is actually based on data-reductionism. It says that the only key to identifying the identity of the self is the flow of data. In other words, non-material consciousness or material body cannot guarantee the identity of the self in a hyperconnected society, and only information can reveal the self. This is why it is predicted that multiple selves will own various identities in the future.

Along with that, some people do not welcome the digitalization of their egos but see it as inevitable (in the case of Kurzweil), while others welcome it and actively pursue it. For example, posthumanists, who have been dissatisfied with the modern view of view, argue that the postmodern self may cause a loss of individuality, but unlike before, it will eliminate discrimination due to differences in consciousness or body and place all members in the net in equal positions.[8]

However, as discussed earlier, we can face a serious moral crisis when the concept of individuality is completely invalidated. Modern culture moved the source of knowledge and values outside the individual (e.g., Plato's *Idea*, *God* in the Middle Ages) to the inside, creating a concept of self as an autonomous moral subject and discovering its value (dignity of human being with self). These discoveries have made it possible for all of us today to be equally respected. In other words, individuality is a philosophical concept established by modern scholars as they newly illuminate individual rational abilities and values. However, if the concept of individuality is completely dismantled, the foundation of human dignity will disappear in an instant. So, what will we now be guaranteed dignity from? In addition, when any moral error occurs, who will be held responsible? And how can each of us manage our lives consistently if we can't clearly identify ***who I am***?

5.2 Suggestion (2) Modifying Modernity

Given the previous discussions, I think that in a hyper-connected society, we should come up with ethical alternatives that can guide individual selves to establish their true identity without going through confusion. This alternative means an ethical culture that emphasizes the concept of individuality that is a base of the concepts of moral responsibility, human dignity, and individual existence, and affirms its value. This culture clearly differs from Nietzsche's dismantlementism. Nietzsche's superhuman ideas is the philosophical root of posthumanism. Posthumanism means overcoming modern humanism, which has led to numerous discrimination and exclusion. However, F. W. Nietzsche (1844~1900) aims to overthrow of all values. Similarly, a lot of posthumanists often argue for a perfect breakaway from tradition (i.e., modernity).

However, I never aim for such a breakaway. Rather, I try to improve by critically embracing tradition. This is because I believe that truth is finally acquired through a process of constantly modifying it through trial and error for many years. In other words, I understand the concept of 'individuality' that emerged after modern times as an intellectual achievement that mankind has expanded through trial and error. We can see from past history how much pain a totalitarian society that ignores individuality can inflict on a human being. In this respect, the recognition of individuality can be seen as a discovery of mankind that can never be discarded so as not to commit another mistake. So, I think there is definitely historicalness and significance to this concept. Therefore, I insist that we should maintain this concept. Then, what is the way to stick to the concept of individuality while overcoming the limitations of modern view of self?

First of all, in my view, the unique identity of the individual self is not realistically fixed, but can constantly change and be multi-layered depending on what environment you live in. For example, we often recognize that my speech and attitude at home and at work change greatly.

For this reason, what my family thinks about me and what my colleague thinks about me may be different. In other words, our identity can be formed differently depending on which group we work in. In this regard, even before transitioning to a hyper-connected society, our identity already can be said to have a multi-layered part. But even if there is such a multilayered aspect, our identity does not remain disorganized in itself. Rather, our identity can be understood and identified more collectively by one subjective consciousness. In other words, the unique single identity that encompasses various aspects of myself can **be *revealed*** in detail by the subjective consciousness contained in one body. However, the ability of reason to identify such an unique identity needs to be described differently from the previous one. Here I'm going to get some help from the modern philosopher, Charles Taylor(1931~), a modern philosopher.

According to Taylor, human reason has the characteristics of 'contextualized autonomy' and 'substantive rationality'.[9] He points out that the real world we live in is not an explicit object. The world is a complex and obscure context and flow. So we can't grasp the world we're in contact with at once. Therefore, he explains that we should materialize the point by consciously clarifying the context and flow around us. In other words, when each of us specifically describes and expresses the world in language, the world reveals its form. Taylor says reflection is necessary to describe and express things. In other words, his reflection is the work of 'self-interpretation'.[10] Through these works, we can evaluate the good or the bad, the high or the low of something. In short, we can find and do what is more appropriate and reasonable in some situations through self-interpretation work. Taylor claims that this is what our reason does. Therefore, Taylor's reason, which is based on dignity, cannot be said to be a simple cognitive or computational ability of objects.[11]

Similarly, I believe that even when we want to identify who we really are, we must perform these specific descriptions and representations. Our identity can't basically be reduced to matter or data (no matter how much data I've produced, if they don't have any meaning to them, they're not worth anything to me). Our identity can only ***be revealed*** when we perform very complex tasks such as interpreting and evaluating the meaning of the object by myself while interacting with the object facing our reason, and describing and expressing the object in detail. Although it is expressed as being revealed here, in fact, the unique self-identity (individuality) is close to what the self, the subject of existence, intentionally ***creates*** by reflecting on one's inner self.

Therefore, it is necessary to awaken the importance of so-called 'self-interpretation' to cyborgs who will live in a hyper-connected society. This is because the existence of 'I' is never a simple sum of data. Our identity or individuality can inevitably be confirmed in the efforts of our own conscious clarification and self-expression. This work can never be done by AI e modeled after the concept of modern reason. This is because today's artificial intelligence is only a computer built from the beginning by imitating only mechanical cognition, calculation, and procedural reasoning, which are part of this performance. So, no matter how much cyborgs live a fast and efficient life combined with AI, their own identity cannot be confirmed only depending on AI. Therefore, it is necessary to emphasize the importance of individuality to cyborgs in a hyper-connected society, and further convey the importance of work that enable them to be identified. Furthermore, it is necessary to remind them how to do such work.

6. Conclusion

The purpose of this study is to analyze ethical problems that may arise in an individual's self-identity against the backdrop of a future society brought by advanced science and technology and to suggest alternatives. The development of AI, sensors, network technology, big data, and computer cloud systems will connect everything, including people, objects, and space, to each other through computer networks in the future. Accordingly, it is expected that information on all things will be produced, collected, shared, and utilized. That is, it is predicted that a hyper-connected society in which things-things, human-things, and human-human connections will emerge, and in this society, the boundaries between the real world and cyberspace, body and machine, self (I) and others will be blurred. And when defining cyborgs as organisms with various technologies, it can be predicted that future humans will naturally become cyborgs fused with various technologies. In particular, I predict that it will be a cyborg connected to an automated machine called AI in the future.

For this reason, we may worry that in the future, new concepts of self that conflict with the concepts of ‘self as a fixed and unified entity’ and ‘self with a unique inner nature’ that are familiar to us will appear, or the concept of ‘individual self’ will disappear. This concern stems from the fact that the concept of independence cannot be established if the concept of self as an individual entity is eliminated. If the concept of individuality is nullified, the following ethical problems can arise:

(1) The question of responsibility (*If you can't tell whose thoughts and actions are mixed up with each other, who will you hold responsible for when a moral error occurs?*

(2) The question of dignity (*how can human values be guaranteed without the concept of individuality? In other words, on what basis will members of a hyperconnected society be respected for their dignity?*)

(3) the problem of existence (*the meaning and purpose of life (moral identity) be identified while individual identity is unclear? How is that possible?*)

We could face serious moral turmoil if these problems actually arise. Thus, we should not abandon the concept of individuality completely. The concept of individuality can be seen as the intellectual heritage of mankind established through trial and error for a long time. Therefore, by maintaining the concept of individuality, it is necessary to help cyborgs maintain their identity consistently and establish themselves properly in the future. In this regard, I would like to present a modified version of the modern view of self. In short, this revised version is a view that emphasizes the special ability of the self that the exiting version overlooked. The modern view of self is understood that the self is a fixed and unified entity and has its own inner self. In addition to it, I would like to evoke that we have the ability of reason (i.e., substantive rationality and contextualized autonomy) to interpret and reflect on our own unique purpose and direction of life by integrating multiple identities formed by interacting others (and the world). This differs from the ability to infer and calculate mechanically. When we recognize these abilities, we humans can finally find an answer to the question, ***how to live?***

In the era of science and technology, which develops day by day, an ethical culture should be formed so that cyborgs can establish their identity properly. However, ethical culture cannot be formed and spread in a day or two. It takes a certain amount of time to form an ethical culture. Therefore, a series of long-term efforts are required to form and spread this culture. this study evokes the importance of the formation and spread of this particular culture. the next study will discuss specific ways to form and spread this culture.

References

1. Gray, Chris Hables. *Cyborg citizen: Politics in the posthuman age*. Routledge(2000): 11.
2. Jordan, John M. *Robots*. MIT Press(2016),
3. https://ahdictionary.com/word/search.html?q=robot.
4. Harari, Yuval Noah. *Homo Deus: A brief history of tomorrow*. random house(2016).
5. Kim, Sun-hee. *Science and Technology and Human Identity*. Acanet(2013): 99-107.
6. Kurzweil, Ray. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. 7. Penguin USA(2000): 490-491.
7. Braidotti, R. *The Posthuman*. Polity (2013).
8. Poster, Mark. "Postmodern Virtualities". in: Meenakshi Gigi Durham. *Media and Cultural Studies*. Wiley-Blackwell (2009): 533-548.
9. Lehman, G. "The Relevance of Charles Taylor's Interpretivism for Environmental Politics" . the Australia New Zealand Society for Ecological Economics 2013 Conference. The University of Canberra and Australia New Zealand Society for Ecological Economics(2014).
10. Taylor, C. *Sources of the Self: The Making of the Modern Identity*. Cambridge, Mass.: Harvard University Press (1989): 72-75.
11. Taylor, C. *Sources of the Self: The Making of the Modern Identity*. Cambridge, Mass.: Harvard University Press (1989): 74-75.

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.010

ETHICAL DISCUSSION ON THE IDENTITY OF DIGITAL HUMAN IN THE METAVERSE ERA

BANG EUN-CHAN

Major of Artificial Intelligence Humanities Convergence, Seoul National University of Education, Seoul, Republic of Korea

E-mail: sbec1210@gmail.com

BYUN SUN-YONG*

Department of Ethics Education, Seoul National University of Education, Seoul, Republic of Korea

E-mail: bsyethos@snue.ac.kr

As the metaverse, recognized for its potential in various fields, grows, the technologies associated with it also develop. Therefore, this study intends to explore the ethical implications of Digital Human technology, one of the core technologies of the metaverse era. First, it examines how the concept of Digital Human is being established, what are the research trends, and how it relates to the metaverse. Second, the essential characteristics of the Digital Human are divided into four categories: 1) Digital Double, 2) Digital Single, 3) Digital Subject, and 4) Digital Object, according to ground and identity, and based on facts and non-facts. In addition, the ADVISE model is proposed as an ethical assessment model based on six major ethical standards (-falsity, discrimination, violence, invasion of privacy, sexuality, and dependence) in relation to Digital Human.

1. Introduction

Facebook, one of the companies that led the golden age of social media platforms at the beginning of the 21st century, changed its name from 'Facebook' to 'Meta' at the end of October 2021. In this regard, Mark Zuckerberg, CEO of Meta, expressed strong anticipation and special confidence in Metaverse in his keynote speech at the company's major annual conference, 'Facebook Connect 2021'. In particular, Meta is currently building the metaverse platform ‘Horizon’, which consists of three main spaces. Moreover, such an image transformation of a socially-led company has become an amplifier of interest in the 'Metaverse' along with related industries.

This expansion of the metaverse market is not simply a monopoly of a single company, but an innovative movement flowing at the bottom of the high-tech ICT industry. Furthermore, a huge impact affects a whole society, including economy, politics, and education, by exceeding the boundaries of industrial fields. Mainly, representative platforms such as Roblox, Zepeto, and ifland operate events, for example fan signing events for singers, conferences and forums of government agencies and companies, and the scope of application is expected to gradually expand.

Among other things, the expandability of Metaverse is maximized by the application of artificial intelligence (AI) technology to realize its potential. In particular, the main focus is Digital Human or virtual human using technologies such as deep learning or communication AI. It seems that interest in robots, AI, and metaverse boils down to 'how can a human or the human world be most realistically implemented'. ‘Oh Rozy’, a fictional character of ‘Sidus Studio X’, a content creative company, has already generated about 1 billion won in revenue in 2021 alone. Also, a number of virtual human cases such as ‘Sua’ from Onmind and girl group ‘Eternity’

* Corresponding author

from Pulse9 have recently appeared. This AI-based Digital Human technology can be used as an avatar of the metaverse, so it can greatly contribute to the detailed realization of the metaverse.

Therefore, this study explores the possibility of a democratic expansion of the Digital Human. As this point, a question arises as to how to assess the identity of Digital Human located between humans and machines. Here, classifying them according to whether they are true or not takes identity and ground as the main reference points. Additionally, when diagnosing the democratic applicability of Digital Human, an ethical assessment model is presented so that the limitations and tasks can be analyzed together. After introducing the discussion of identity and ethical considerations, it predicts what role Digital Human can play in a democratic society as they advance technologically in the future. In addition, it is intended to lay the foundation for the ethical development of Digital Human by examining the ethical issues and proposing future tasks.

2. Digital Human Identity Discussion

2.1. *Concept and Development Direction of Digital Human*

As Homo Faber, the instrumental and technological expertise of human being has now focused its attention on the representation of human being or their world, starting with the primary use of natural materials. The concept of Homo sapiens, which had led Homo Faber, is no longer leading the production of its meaning, but rather a reversal that the meaning of Homo sapiens is defined by Homo Faber is occurring. The future of modern ICT technology is condensed into Reality (R), which is the limit of representation. This is revealed by the development of reality-based technologies such as VR (Virtual Reality), AR (Augmented Reality), DR (Diminished Reality), MR (Mixed Reality), and XR (eXtended Reality). Additionally, reality digital devices such as VR HMDs like Meta’s Oculus Quest and Microsoft's HoloLens, and AR devices are also developing as a means to support the reality-based technologies. Accordingly, the provisional layers or areas (of multiple mixed reality and virtual) that exist infinitely between the perfect reality and the perfect virtuality exist variously and overlappingly between Augmented Virtuality (AV) and Augmented Reality. [Figure 1] shows the spectrum of such virtual and reality. Virtuality can be divided into surrealistic virtuality and realistic virtuality, then AV comes from the latter. The overlapping phenomenon between reality and virtuality intensifies in Hyperreality where AV, which adds reality to realistic virtuality, pursues extreme reality, or where AR, which adds realism to reality, becomes overly realistic. In conclusion, virtuality and reality, which were located in opposite concepts, are each augmented and connected as one, forming a new world. In this paper, this is called 'HYPERREALITY'.

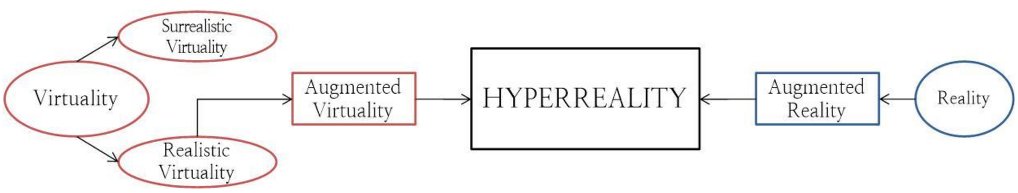


Figure 1. The spectrum of virtuality and real.

Meanwhile, a key issue in the process of world construction in virtual space is how to implement a virtual human closer to a more perfect reality by actively utilizing the developing technologies. In an issue report of the Software Policy Research Institute, Seung-Hwan Lee and Sang-Yeol Han predict the future through the spread of Metaverse and identify the five major issues related to Metaverse as "① Beyond Game, Toward Economy ② Expanding Metaverse Devices ③ Growing Digital Human ④ IP x Metaverse ⑤ NFT x Metaverse" [1]. When discussing the growth of Digital Human among the five issues, they revealed that Digital Human is being used in various fields as "technological constraints are relaxed" and AI is applied [1]. This rapidly emerging Digital Human implementation technology is expected to make a important contribution to the final position of the metaverse, a virtual expansion of the human world.

The concept of Digital Human is still being actively studied, and only the initial concept in the company has been established, and related research is still in its early stages. Looking at the discussions of companies, first, Virtuals, a virtual production platform, outlined Digital Human as "a photorealistic 3D human model" [2]. They characteristically used a method of conceptualizing through comparison with several similar concepts. They summarized the distinction between a general 3D human or a deepfake and Digital Human as a "structured 3D model" [2]. As a result, only Digital Human can be controlled, and it is confirmed that it is a different concept from 2D-based deepfake. Because of the technical flexibility of the digital human due to this controllability, Digital Human can be used in various situations. And as a sub-concept of digital human, "virtual human" and "digital double" are divided. First, virtual human refers to people who have their own unique (even if created) identity to some extent, such as Sidus Studio X's Oh Rozy and the first virtual influencer Lil Michela. Second, digital double is a copy of a real person, unlike virtual human, and thus there is a risk of legal or ethical problems. In any case, their Digital Human conceptualization model is meaningful in that it separates similar words and suggests inclusion relationships.

In some cases, AI is expected as the core technology needed by Digital Human. Saltlux, an artificial intelligence and big data technology company, defined Digital Human as "a 3D virtual human who looks like a real person and speaks like a real person" [3]. In addition, he added, "AI-based, lifelike character" through the term "AI human" [3]. This shows that the application of AI in the technological development of Digital Human is the best way to increase the human imitation rate of Digital Human and is the latest trend. Deloitte, a global accounting firm, conceptualizes the Digital Human as "an avatar that can produce a whole range of human body language" [4]. This discussion is equivalent to defining Digital Human with a focus on communication. They assumed that AI-based Digital Human have strengths in verbal communication by using interactive AI for Digital Human, and non-verbal communication based on AI's learning abilities such as facial expressions and hand gestures.

In spite of the lack of discussion on Digital Human in academia, there are some studies that are linked to production techniques based on the analysis of the opinions of leading companies. First, Seung-Hwan Lee and Sang-Yeol Han presented Digital Human as "3D virtual human with a shape similar to human appearance/behavior" [1]. In fact, Digital Human development company UneeQ has developed a customizable UneeQ Creator based on its own designed Digital Human, and the generalization and facilitation of Digital Human production is progressing. A study, which analyzed the production techniques of Digital Human and metahuman, said that the definitions of Digital Human mentioned above are established in a way to "①imitate the appearance of a real person, ②imitate the behavior of a real person, and ③replace the role of a person" [5]. They defined Digital Human as "a 3D human body model created by imitating the characteristics and appearance of a real person for the purpose of replacing human roles" based on commonalities [5]. Here, the purpose of the Digital Human as a 'human substitute' was confirmed. Also, they especially looked at the emergence of 'metahuman' as a progressive concept of Digital Human. It was borrowed from a metahuman creator using Epic Games' game engine, Unreal Engine. At the same time, they said that the previous method of the Digital Human was a ‘passive virtual human’ requiring manual labour, but the meta-human was an ‘active virtual human’ that is a mixture of a virtual human and a digital double.

As a result of some of the widely used concepts of Digital Human, they show that Digital Human, a relatively new ICT technology, is not yet organized or converged as the concepts of MR and XR, Digital Human, and the research is actively being conducted. Among them, based on the definition of Oh Moon-Seok et al., who identified relatively common characteristics, this study defines that Digital Human imitated human using digital technology to replace humans.

2.2. *Digital Humanism: Between Facts and Reality*

The ontological investigation of human identity, including objects that share similar or common characteristics with human, continues in various technological, material-ideological, and technical-ideological forms such as Cyborg, Neo-Humanism, Post-Humanism, Transhumanism. Excluding Neo-Humanism, which tried to philosophically reclaim the meaning of human existence under the assessment that human beings are facing an existential and ontological crisis [6], most of them can be seen as belonging to the technology-oriented Modern Humanism trend.

However, in the current of modern humanism ideology ontology about humans in the real world such as transhumanism has been mainstream in the academic and technical discourse for reasons of reality and threat. Perhaps it is because the avatars in the game, which have maintained their digital object, have significantly reduced their similarity to humans, and their development direction has been almost limited to the entertainment market. However, with the advent of the metaverse 2.0 era, a new request for digital humanism is being raised in such a humanistic discourse. Given the diversification and advancement of Digital Human, there is an urgent need to reinterpret Digital Humanism through comparison with a human being like other humanism discourses. Through this, the identity of Digital Human will be established and ethical and legal discussions about Digital Human will be more sophisticated.

When discussing the identity of a digital human, the ‘virtuality’ of the digital space should be considered first. This can be divided into types in which ground and identity meet in the relationship of fact and non-fact, respectively. The first is ground, based on which of the facts and non-facts the Digital Human is produced. In this case, the division is relatively easy. First, there is an example where Digital Human, based on facts, has the intention to copy and clone a real person and realize it as precisely as possible. It can be said that the ‘Digital Double’ of Virtuals as a name and the ‘Human Digital Twin (HDT)’ or ‘Digital Twins of the People (DToP)’ as a derivative of the industry’s digital twin technology are applicable. This paper will summarize it as a ‘Digital Double’, a digital doppelganger reproduced in a parallel world where only signals of 0 and 1 exist. An example of Turtle's leader, TurtleMan, restored in the AI music project 'Once Again', was completed based on two facts, the singer's voice and face. At this time, the AI learned the voice data from the actual singer's audio data and sheet music data. Also, AI learned facial data by using face editing technology that specialized in generative adversarial neural network (GAN) technology for faces. Other examples include the reincarnation of Carrie Fisher, who died of heart disease in 2016, as Leia in the Star Wars series, and the survivor ABBA's virtual concert scheduled for 2022. Currently, the digital double focuses on the digital revival of the deceased. Nevertheless, it has infinite potential for expansion in terms of use, such as the 20th presidential campaign in which AI presidential candidates appeared, and metaverse-based work or learning that participates as a virtual entity.

Next, Digital Human based on non-fact can be regarded as a digital object created with the purpose of creating an independent virtual human-like object. The name is mainly 'virtual human', and everyone, including real humans, becomes a virtual human from the point of view of all digital objects as virtual. Therefore, in contrast to the digital double, it is summarized as ‘Digital Single’, which is realized as the only human in the digital world. One example of this is the virtual influencers that have been rampant in the marketing world recently. AYAYI and Liu Yexi from China, also called virtual KOL (Key Opinion Leaders), Oh Rozy and Ms. XEN from Korea, and Lil Michela and Shudu are those cases. In addition, when Digital Single is used industrially, it is expected to have big potential as AI show hosts, VR game avatars, AI teachers, etc.

The second position on virtuality is a discussion of facts and non-facts about identity. Here, it seems to be somewhere on the continuum between fact and non-fact through a complex thought process according to philosophical views on existence and non-existence, real and fake, life and non-life, and humanistic discourse. In particular, since this study objectified Digital Human above all else, the discussion was focused on the possession of the Digital Human's own identity.

First, distinguishing the identity of Digital Human as a fact is equivalent to acknowledging its own existence, so it is called a ‘Digital Subject’ by acknowledging its subjectivity. The modern French sociologist Jean Baudrillard explained simulacre as an example of Mickey Mouse. He saw that Mickey Mouse, cloned from the reality of the mouse, is not regarded by us as a clone of the mouse (even disconnected from the reality of the mouse), but as an independent sign [7]. Therefore, Digital Human that has characteristically cloned a human is also an independent and a truth independent of the reality of human being. In a place where even reality is considered virtual and even digital networks are connected to reality, it is difficult to even say that Digital Human are not real [8]. As an example of standing at the center of this philosophical confusion, virtual influencers appear who are considered to already have their own identities such as worldview and inclination. They have with age and gender from the beginning and are born with hobbies and occupations, and their identities is recognized, and they even work as advertising models and ambassadors. Then, the answer to the question of what is different from the existing dolls or photos is the possibility of ‘owning reason’ due to the application of AI, and the fundamental root is the same as that of the Stoic school, which established the universality of human being from reason [9]. Of course, it is clear that AI is still far from being compared to human consciousness, so it cannot be perceived as a homogeneous identity with human. Nevertheless, they were considered Digital Subject because they had an identity as Digital Human themselves.

Next, it is a case where the identity of Digital Human is belonging to the realm of non-fact, and the technical and material properties of the digital person are assumed to predominate. This can be called a 'Digital Object' as it rejects the independent and subjective ability of Digital Human. By denying the identity of Digital Human, it is intended to give only a narrow role as a puppet and to define it as an instrumental character. This is merely diagnosing the Digital Human as a technologically progressive aspect of the existing game avatar. One such example is Parzival in Oasis in the movie 'Ready Player One' or Neo in the Matrix in 'The Matrix'. Digital Human can be said to be Digital Object when it is possible to give a negative answer to the question of whether it has an independent identity in terms of the identity of Digital Human. They are just like ‘shells of the soul’ that fill the human identity.

<Table 1> shows the virtuality dimension of this Digital Humanism by setting the x-axis as ground and the y-axis as identity. Ground, one of the two criteria, depends on whether a real person is cloned. Another criterion, identity, depends on whether Digital Human possesses a

unique identity for itself. Next, Digital Human can be divided into four areas according to two criteria. First, Digital Single-Subject is an independent human analog entity that exists only in the virtual world but is Digital Human with a unique identity. A prime example of this is the virtual influencer. Second, Digital Double-Subject possesses a unique identity and is Digital Human replicating a real person as it is, and Digitial Double even recreate deceased people. Third, Digital Single-Object is a being who lends an identity like a character in a game, and it means Digital Human who has neither his own identity nor a replica model. Examples include Parzival in Oasis in the movie 'Ready Player One' or Rui, a virtual YouTuber in dob studio. Fourth, Digital Double-Object is Digital Human that exists as a soulless replica and agent of a real person without a unique identity, such as ‘Neo’ in the matrix of the movie 'The Matrix'.

Table 1. Classification of Essential Characteristics of Digtital Human

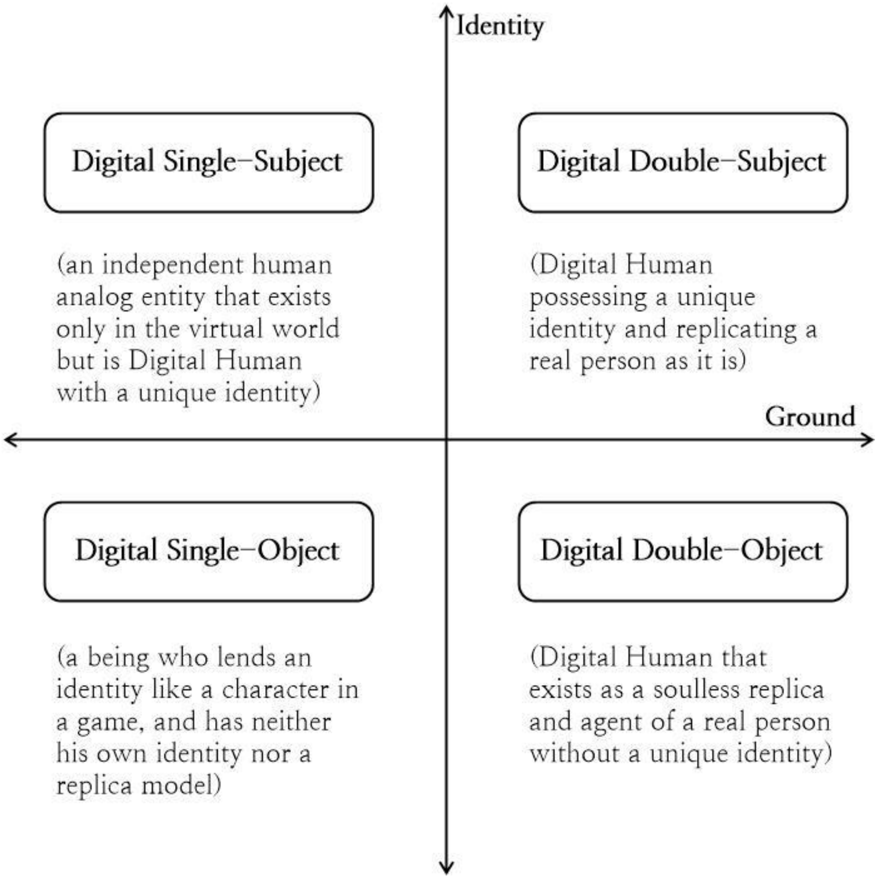
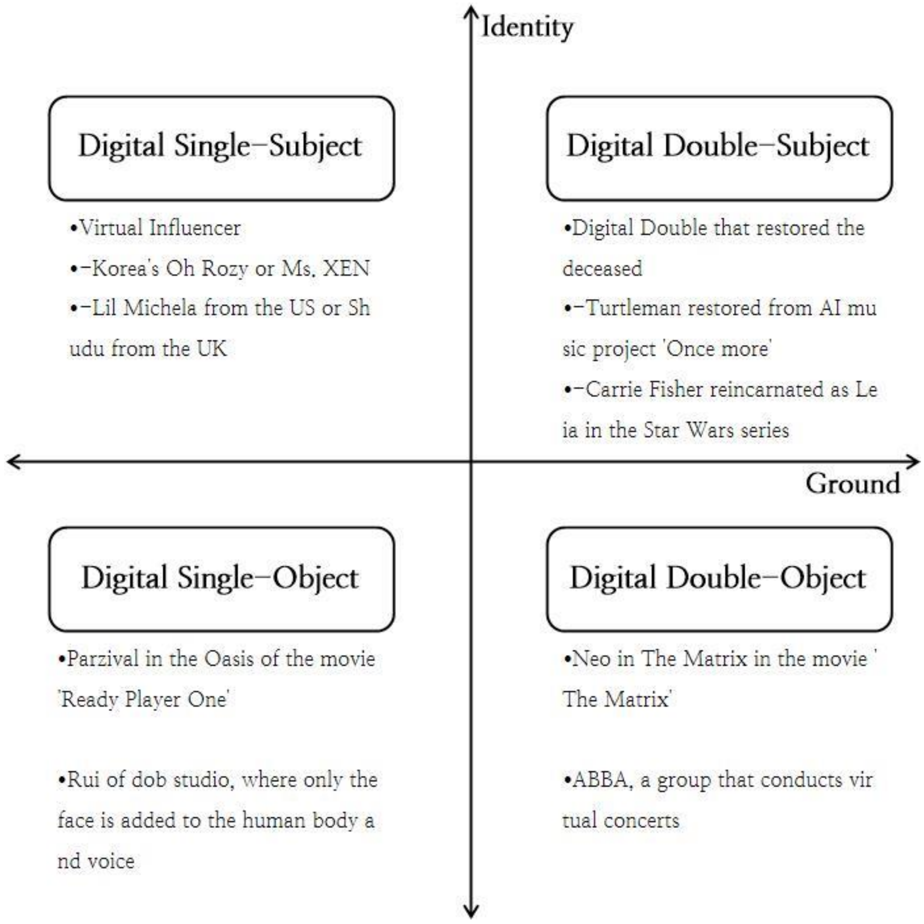


Table 2. Cases of the Essential Characteristics of Digtital Human



3. ‘ADVISE’, Assessment Model for Ethical Responses of Digital Human

Although there are still many issues to define the identity of a digital human, an ethical problem for digital humans is emerging as the content unethical of digital objects (-non-facts). Digital Object is a being in which meaning is given only when it is used and conscious of by a human, and human desires are inherently behind all digital virtualization [10, 11]. Ultimately, the ethical or unethical of a digital human looks like a problem that depends on the desire or intention of the person who uses it. Therefore, in the unethical contents of the digital human, the person must actively play the role as the subject of responsibility. Responsibility here entails not only causal responsibility in the traditional ethical conception, but also an extension to predictive responsibility, as Hans Jonas argues [12]. Also, due to their similarity to humans, the possibility of human rights violations is high when Digital Human is used unethically. A sense of ethical responsibility is required for each actor, such as developers and manufacturers, public entities, users, and educational entities. However, ethical measures instead of legal and technical measures first require an amicable social consensus among the subjects. For this purpose, the ‘ADVISE (fAlsity, Discrimination, Violence, Invasion of privacy, Sexuality, dEependence)’ model is presented as an assessment model for ethical response as shown in [Figure 2]. In the ADVISE model, the defects of various digital technologies are comprehensively considered due to the convergence technology aspect of Digital Human.

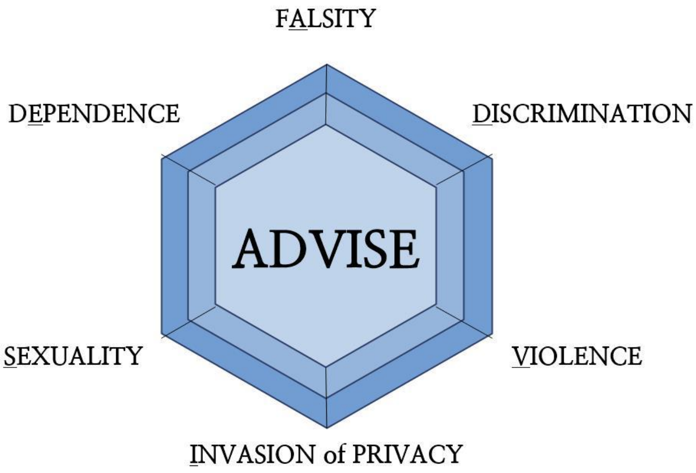


Figure 2. Digital Human's ethical response assessment model.

The first is fAlsity. Falsity is a prominent ethical situation in the Digital Double, and it is a question of how to protect the Digital Double's existential model. The purpose of the Digital Double is to be a perfect replica of a real person, so an explicit or implicit agreement with the target is required. In the case of photo that is 2D-based image human, even if they are real or falsity in a way, they are protected by the right of portrait, and if they are taken without consent, they must bear legal punishment as well as ethical responsibility. Therefore, Digital Double are also 'falsity of real', but a clear agreement on protection should be drawn up on the same line as photos. However, all Digital Double, except for the agreement situation, should be classified as 'falsity' in reality. Nevertheless, ethical requests for the falsity of this Digital Doubles are insufficient apart from for the awareness of sexual objectification. Even looking only at the cases of deepfakes, no specific guidelines were given, other than those stipulated in the Sexual Violence Punishment Law, which was amended in 2020. Also, the perception of falsity creation is revealed from the standpoint that it is different from falsity because the texts and pictures that sexually objectify male idols are only imaginary just by looking at the controversy caused by the RPS (Real Person Slash) Punishment Act promoted by some lawmakers. Despite the rejection of the position on both cases, however, the creation of Digital Double is very realistic, so stricter ethics against falsity seem important. In addition, in the case of deepfake, a transitional technology of digital double, the negative contribution to phishing and disinformation is increasing [13]. Therefore, appropriate ethical assessment and social consensus are required according to the trend of deepfake abuse cases.

The second is Discrimination. Bias in the world of artificial intelligence, discrimination and damage caused by it are the most discussed ethical issues. Most notable has been the discriminatory use of textual language by chatbot AI such as Tei and Iruda, but now, examples of discrimination related to images often emerge. A typical example is the Google Photos incident in 2015, which caused controversy because black people were classified as gorillas [14]. In 2020, Twitter apologized for racism in favor of white people over black people in its image cropping algorithm [15]. Seo et al. mentioned artificial intelligence as one of the essential technologies for precise Digital Human production [16], and the degree of technological intervention of artificial intelligence in Digital Human is rapidly increasing. Among them, the learning process for images is also included, so the case of discrimination of artificial intelligence above is sufficiently applicable. Also, they say that artificial intelligence can be effectively used for interaction through natural language input and output in addition to image learning.

The third is Violence. The two major mountain ranges of diagnosing social dangers from digital media were sexuality followed by violence, and several studies analyzed both scales together [17-19]. Studies on whether such digital violence has a real social impact show both pros and cons of its causality [18, 20, 21]. However, the digital world such as the metaverse in which Digital Human acts is a comprehensive society that differs only in the signal system from the real world. Therefore, in the digital world, violence within the virtual world should be deal with at a short-distance manner, unlike the long-distance approach to violence in existing digital media or games. In this regard, the International Center for Countering Digital Hate (CCDH), a non-profit organization, reported a study result showing that one case of sexual abuse is exposed in Metaverse every 7 minutes [22]. Moving away from the traditional view of violence in the digital world seems to be the key, such as analyzes discussing the mixed nature of the metaverse and the real [23] or the differences between the metaverse and other internet platforms as a possibility of general action [24].

The fourth is Invasion of Privacy. The precious value of data is highly emphasized in the intelligent information society to the extent that data in the information revolution is given a position corresponding to the fossil fuel in the industrial revolution [25]. Contrary to ethical desires, sexual crimes based on the covert sex industry or unhealthy sexual perception grow together with technological advancement. After entering the digital society, conceptualization of online and cyber sexual violence began around the 2000s in Korea, and online sexual violence diversified with the diversification of media and platforms due to the development of digital technology [26]. Afterwards, the aspect of the development of sexual crimes was confirmed according to the technological leap, on the reconceptualization of non-consensual videos and image-based sexual exploitation along with reports of revenge porn around 2007, when smartphones and mobile technologies were developed. Therefore, it is highly likely that both Digital Single and Digital Double will become embroiled in controversy over sexuality, and ethical discourse should be raised accordingly. This is true whether in the form of Digital Human avatars or algorithmically controlled Digital Human agents, which they are the two classes of virtual human [27, 28]. For example, if an AI agent or AI anchor wears revealing clothing, an issue will arise as to whether it is inappropriate, and the extent of touch or sexual behavior within the metaverse is also controversial. In fact, in December 2021, the meta announced that it would introduce a "personal boundary" in the company metaverse as a technical measure to the case of physical contact between avatars reported in December 2021. Since an increase in related cases is inevitable in the future, a deep deliberation on the sexuality of Digital Human seems urgent.

The sixth is Dependence. In general, AI development types are divided into weak AI, strong AI, and super AI. In most cases, it is considered that it will take more effort before the arrival of strong artificial intelligence that is capable of many things like human [29-31]. The future value of Digital Human has a high proportion of collusion with artificial intelligence and will follow such a growth line, so the problem of dependence and reliability arises in the gap. In this 'dependence gap', ethical questions continue to be raised about how far to depend on the role of Digital Human, which extends to the issue of reliability. Looking at the convergence of Digital Human and politic, it seems that the AI presidential candidates that appeared in the 20th presidential election are not yet dependable Digital Human. Also, Digital Human that act as advisors to policymakers, such as AI assistants, are caught by general doubts about the effectiveness of AI decision-making [32], just like in the Japanese mayoral election in Tamashi [33]. Even other studies on the effectiveness of AI decision-making have given a yellow signal to the rise of an AI politician or secretary, indicating that reliance on AI decision-making can make negative choices [34-38]. In other words, if the stable technology of artificial intelligence is secured, it strongly proves that Digital Human can be used in "mediated politics" mediated by an AI virtual assistant that appears in future political scenarios according to the trend of artificial intelligence by Son [31]. Here, there will be a difference in dependence between "mediated politics" and a robot government [31] where humans are eaten by super artificial intelligence. Therefore, it is necessary to stably lead the growth process of the Digital Human, a transitional being, through a predictive ethical perspective on dependence.

4. Conclusion

Despite the uncertainty of the future society expected from the development of current technology, the responsibility given to us now is to prepare for changes in our perceptions and attitudes. In this way, the technological changes that are approaching us can have a positive impact on the happiness of human life as much as possible. Thus, it becomes very important how one perceives and reacts to the emergence of the metaverse and related technologies in the age of intelligent information, which is immediately imminent. Accordingly, this study has considered the Digital Human presented as a part of the core technology of the metaverse from an ethical and political perspective.

The Digital Human poses a new characteristic problem about the essence of human beings due to its fundamental characteristics that cannot be included in the modern concept of Humanity. The identity of Digital Human beings is located between facts and non-facts. At present, the Digital Human seems to be close to the virtual, but there may come a time when it comes close to the truth, so a clear perspective on identity is needed. Therefore, in this study, Digital Human was divided into four types according to the identity and ground criterion. After combining all four categories, identity classification was again presented in four categories. The reason for this discussion is important is that the ethical approach to secure the safety of technology is different for each identity. In addition, the 'ADVISE' model was presented as a assessment model to contribute to ethical stability. This is to enable ethical filtering of major issues according to the development of Digital Human, but it lists rather narrow ethical standards, which is seen as a limitation. Therefore, the process of continuously tracking, correcting, and supplementing ethical issues related to Digital Human should be accompanied in the future. However, as this is still an early-stage study, an ethical response according to the improvement of digital human technology will be conducted as a follow-up study.

References

1. S. H. Lee and S. Y. Han, *Metaverse begins : 5 Major Issues and Forecast***(Report NO. IS-166)**. Gyeonggi-do: Soft Policy Research Institute, (2021).
2. L. T. Alvaro, *Digital Human, virtual humans, digital doubles ... what's the difference?*. **Virtuals**, (February 12, 2020). Digital Human, virtual humans... what's the difference? | Virtuals (visited February 26, 2022).
3. *The value and usefulness of a 'Digital Human', the protagonist of the metaverse era*. **Saltlux**, (March 9, 2021). https://blog.naver.com/saltluxmarketing/222269505818 (visited February 28, 2022).
4. *Digital Human, Elevating the Digital Human experience.* **Deloitte**, (n. d.). https://www2.deloitte.com/nl/nl/pages/customer-and-marketing/articles/digital-human.html (visited February 26, 2022).
5. M. S. Oh, G. H. Han and Y. H. Seo, *A Study on the Production Techniques of Digital Human and Metahuman for Metaverse*. **Korea Institute of Design Research Society, 6**(3), 133-142 (2021).

6. J. K. Lee, *On Confrontation and Reconciliation between Technofuturism and Neohumanism*. **Phenomenology and Modern Philosophy, 59**(-), 5-49 (2013).
7. E. G. Cho and D. R. Chang, *A Study on the Luxury Brands' SNS Content Modern Artification-Fousing on Jean Baudrillard's Simulacra-*. **A Journal of Brand Design Association of Korea, 19**(3), 235-252 (2021).
8. H. T. Kim, *Virtual and Real, Human Life and Significance in Posthuman Era*. **Semiotic Inquiry, 59(-),** 89-112 (2019).
9. S. R. Kang, *Stoicism and Cosmopolitanism*. **Journal of Education for International Understanding, 14**(2), 39-60 (2019).
10. D. Y. Kim, *In Preparation for the Age of A.I., Apply Hans Jonas' Theory of Responsibility*. **The Journal of Elementary Education, 25**(-), 63-88 (2021).
11. S. H. Kim, *Philosophical Analysis of De-realization with the Smoothness of Digital Virtualization: Focusing on Nietzche and Han Byung-chul*. **Studies in Philosophy East-West, -**(100), 323-350 (2021).
12. K. Jung, *A Critical Review of Hans Jonas' Responsibility Ethics-Focusing on Meta Ethics and Neuroscience Approach-*. **Journal of Ethics, 1**(120), 185-210 (2018).
13. J. I. Park, *The Development of Deepfake Technology and the New Threat of Journalism-the Necessity of Image Fact Checking to Ensure Image Fairness.* **Press arbitration quarterly, 156**(-), 52-69 (2020).
14. B. G. Goo, *The Blind Spot of Machine Learning, What the 'Black = Gorilla' Error Tells Us*. **The Hankyoreh**, (April 5, 2019). https://www.hani.co.kr/arti/science/future/877637.html (visited February 28, 2022)
15. Y. J. Yun, *Twitter apologizes for racial bias in image cropping algorithm*. **AItimes**, (September 22, 2020). http://www.aitimes.com/news/articleView.html?idxno=132361 (visited February 28, 2022)
16. Y. H. Seo, M. S. Seok and G. H. Han, *The present and future of the Digital Human*. **Korea Society Broadcast Engineers Magazine, 26**(4), 72-81 (2021).
17. J. H. Baek, D. K. Lee, C. Y. Hong and B. T. Ahn, *Multimodal approach for blocking obscene and violent contents*. **Journal of Convergence for Information Technology, 7**(6), 113-121 (2017).
18. M. S. Shim, E. H. Park and K. H. Kim, *Sexually Suggestive and Violent Features of TV Programs Revealed in Violations of Broadcasting Standards*. **Korean Journal of Broadcasting and Telecommunication Studies, 25**(5), 159-204 (2011).
19. H. S. Yu, *A Study on the Re-establishment of Broadcasting Journalism Ethics in the Digital Medium Era: Sensational and Violence-latent Content, and Digital Manipulation in Broadcast News*. **Korean Association for Broadcasting & Telecommunication Studies, -**(56), 61-87 (2003).
20. H. R. Lee and E. J. Jeong, *A Study on Critical Debate of Violent Games based on Catharsis theory & Cognitive neoassociation theory*. **Journal of the Korean Society for Computer Game, 27**(3), 181-191 (2014).
21. D. H. Han and Y. S. Lee, *The Effect Internet Video Game on the Aggression in Individuals with Internet Game Play*. **Journal of the Korean Neuropsychiatric Association, 52**(2), 57-66 (2013).
22. *Facebook's Metaverse is unsafe*. **Center for Countering Digital Hate (CCDH)**, (n. d.). https://www.counterhate.com/metaverse (visited February 26, 2022)
23. H. J. Yun, J. Lee and H. Y. Yun, *A Preliminary Study on Concept and Types of Metaverse: Focusing on the Possible World Theory*. **Korea Humanities Content Society, -**(62), 57-81 (2021).
24. H. O. Lee, *Study on the Protection of Fundamental Rights and Platform Regulation in Metaverse*. **Constitutional Law, 27**(4), 127-179 (2021).
25. D. J. Hand, *Aspects of data ethics in a changing world: Where are we now?*. **Big data, 6**(3), 176-190 (2018).

26. S. A. Kim and D. H. Jang, *Online Sexual Violence and Sexual Objectification: through Victimized Experience Online*. **Korean Women's Communication Association, 34**(1), 89-130 (2019).
27. J. Fox, S. J. G. Ahn, J. H. Janssen, L. Yeykelis, K. Y. Segovia and J. N. Bailenson, *Avatars versus agents: a meta-analysis quantifying the effect of agency on social influence.* **Human-Computer Interaction, 30**, 401–432 (2015).
28. G. Suárez, S. Jung, and R. W. Lindeman, *Evaluating virtual human role-players for the practice and development of leadership skills*. **Frontiers in Virtual Reality, 2**, 31 (2021).
29. M. Bialik, C. Fadel and W. Holmes, *Artificial Intelligence in Eeducaiton, Promises and Implications for Teaching and Learning* (translated by J. Y. Jeong and S. B. Lee). **Park Young Story**. (2021). (original work publication 2020)
30. J. S. Park, *Response Strategy on Artificial Intelligence by Major Countries and Proposal for Political Development of Korea*. **Dankook Law Review, 41**(3), 35-73 (2017).
31. H. J. Son, *Artificial Intelligence Revolution and Future Scenarios for Politics*. **Journal of Regional Studies, 28**(2), 61-87 (2020).
32. X. Wang and M. Yin, *Are Explanations Helpful? a Comparative Study of the Effects of Explanations in AI-assisted Decision-making*. **In 26th International Conference on Intelligent User Interfaces**, 318-328 (April 2021).
33. S. U. Yoo, Is an 'AI Politician' without Greed better? … Japan's Local Elections, Artificial Intelligence Running. The JoongAng, (December 7, 2019). https://www.joongang.co.kr/article/23651227#home (visited February 28)
34. G. Bansal, T. Wu, J. Zhou, B. Nushi, E. Kamar, … and D. Weld, *Does the Whole Exceed Its Parts? the Effect of AI Explanations on Complementary Team Performance*. **In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**, 1-16 (May 2021).
35. Z. Bucinca, M. B. Malaya and K. Z. Gajos, *To Trust or to Think: Cognitive Forcing Functions Can Be Misleading in Evaluating Explainable AI Systems*. **In Proceedings of the 25th International Conference on Intelligent User Interfaces**, 454-464 (March 2020).
36. Z. Bucinca, P. Lin, K. Z. Gajos and E. L. *Glassman, Proxy Tasks and Subjective Measures Can Reduce Overreliance on AI in AI-assisted Decision-making*. **Proceedings of the ACM on Human-Computer Interaction, 5**(CSCW1), 1-21 (2021).
37. A. Bussone, S. Stumpf and D. O'Sullivan, *The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems*. **In 2015 International Conference on Healthcare Informatics**, 160-169.
38. M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez and K. Z. Gajos, *How Machine-learning Recommendations Influence Clinician Treatment Selections: the Example of Antidepressant Selection*. **Translational Psychiatry, 11**(1), 1-9 (2021).

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.013

THE MEANING OF 'SMART' IN SMART CITIES AND ETHICAL DEBATES ON SMART CITY

YEONHEE LEE and JIWON SHIM

Educational Institute, Gyeongsang National University, Jinju, South Korea

Department of Philosophy, University of Dongguk, Seoul, South Korea

E-mail: proy13@naver.com, g1dmpkr@gmail.com

www.gnu.ac.kr. www.dongguk.edu

The discussion on smart city has been actively discussed in Korea since the event of AlphaGo, and the establishment of a smart city is contemplated worldwide. However, there are relatively few discussions about the meaning or value of smart cities. The term 'smart' is used indiscriminately in various fields without a clearly defined concept. What exactly is a smart city? In the smart city being discussed today, it is necessary to discuss what it means to be a smart city and what the smart city we envision is aiming for. In this paper, the meaning of smart in smart city and social and ethical issues that will arise when smart city is established are discussed. First, smart cities, which utilize a large amount of information, deepen the "privacy issue" raised by M. Foucault (1926–1984). Second, there is a problem that 'intelligence', which is important in smart cities, is defined in a very narrow sense. Third, smart cities that try to connect everything can give rise to 'totalitarianism'. The value that smart cities are aiming for does not just mean technological intelligence. Also, the idea of why to make a smart city should take precedence in the smart city discussion rather than the idea of which city to envision. What is important in the smart city discussion is the 'city', not the meaning of smart.

1. Introduction

The discussion on smart city has been actively discussed in Korea since the advent of AlphaGo, and the establishment of a smart city is contemplated worldwide. However, there are relatively few discussions about the meaning or value of smart cities. The term 'smart' is used indiscriminately in various fields without a clearly defined concept. What exactly is a smart city? In the smart city being discussed today, it is necessary to discuss what it means to be a smart city and what the smart city we envision is aiming for. In this paper, the meaning of smart in smart city and social and ethical issues that will arise when smart city is established are discussed. First, smart cities, which utilize a large amount of information, deepen the "privacy issue" raised by M. Foucault (1926–1984). Second, there is a problem that 'intelligence', which is important in smart cities, is defined in a very narrow sense. Third, smart cities that try to connect everything can give rise to 'totalitarianism'. The value that smart cities are aiming for does not just mean technological intelligence. Also, the idea of why to make a smart city should take precedence in the smart city discussion rather than the idea of which city to envision.

2. The Meaning of 'Smart' in Smart Cities

The dictionary definition of the term 'smart' is when used along with software or hardware, that it has an ability to process information that it could not be expected so far. Also, the potential value of Smart Products is introduced as elements that are based on the use of information and can provide better functionality or service that can increase the value of relationship-based services. But the term 'smart' is being use much indiscriminately, in the sense that almost all home appliances are being launched with the prefix ‘smart’ and the term is being used as a new paradigm in cities and spaces.

For the ‘smart’ devices such as smart phones, smart bomb, smart card, and smart air conditioner, the term is commonly used with the aim of delivering the concept of intelligent and advanced. The term ‘smart’ used with cities and spaces such as smart grid and smart house tries to show that cities or spaces are eco-friendly and be built with advanced technology.

However, while definitions for some ‘smart’ products exist, it is impossible to find generally accepted and agreed upon definition of smart products in the literature due to its indiscriminate usage of the term in the short term. McFarlane et al.’s definition states that intelligent product should be uniquely identified, and can communicate effectively with its environment, storing and retaining data about itself, deploying a language to display its features and requirements, and finally capable of participating in or making decision relevant to its own destiny[1]. However, majority of the literatures does not provide a definition, demonstrating the need for a consensus definition.

Several problems arise from this conceptual confusion. Without a common base, it would be difficult to build a theory specifically for this system, and any discussion may end with lack of mutual understanding due to the confused use of the term, making the establishment of definition of ‘smart’ products a priority. Without the lack of common definition, smart products could end up as another example of disappearing buzzword.

Especially for the term ‘smart cities,’ governments and research since the 1990s used the term as a fashion label in the hope for certain cities to be distinguished with other existing cities and promoted as innovative without long term plans for achieving the purpose, making the definition of the term ambiguous and making it a challenge for others that are facing the process for developing plans for the aspiration of smart cities.

There were number of studies attempting to define the concept of the Smart City, but due to its multidisciplinary concept and the ambiguity of the term ‘smart,’ the agreed-upon concept of the term does not yet exist.

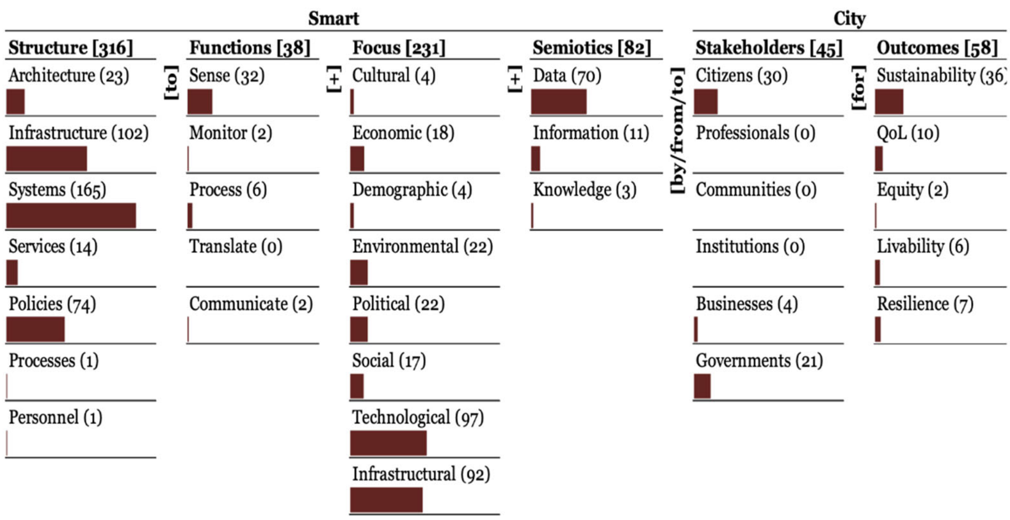


Figure 1. The ontological Map of Monads of Smart City Research.
(Citation: Ontological Review of Smart City Research, A Ramaprasad, AD Sanchez, Thant Syn, 2017)

As in the ontological map above which analyzed 373 articles with the theme of smart city, most of the articles in the corpus mainly focused on ‘smart’ aspect of smart city mostly focuses on technological and infrastructural aspects of smart city management systems, whereas environmental and social aspects receive less focus. In addition, literatures mainly focused on ‘city’ aspect of smart city mainly deals with stakeholders and outcome, focusing sustainability and quality of life, showing the gap of the definition used by various field of research.

The early attempts to define the term were focused on the smartness provided by IT (Information Technology) for managing functions of the city, represented by the concept that IBM has used in 2010: instrumented, interconnected, and intelligent city. Later, the studies have widened their scope to include the outcome of the various concepts such as sustainability and quality of life. Despite of these attempts to describe the term, there are more than thirty-six definitions of smart cities.

3. Ethical discussion about ‘Smart city’

The term 'smart' is used indiscriminately in various areas without the concept clearly defined. Thus, the term, when used to express a new paradigm of a particular space called 'city', does not give a clear indication of what it means. For example, the term 'smart city', which is a commonly used term in urban construction policy and industry at home and abroad, implies the character and vision of future city that we can pursue. However, according to an investigation by the International Telecommunication Union (ITU), there are about 116 definitions of 'smart city'[2] . Different countries (or institutions, companies) use the term in different ways. For this reason, there is a certain difficulty in clearly understanding the nature and direction of the city, which is alluded to as 'smart city'.

Nevertheless, if we look at the urbanization policies of the countries that are currently under construction of 'Smart City', we can roughly examine the nature and direction of future city that is suggested as 'Smart City'. Smart city, as its name suggests, can be summarized as a 'smart city'. The question then might be, "What is a smart city?" Or "How is a city smart?" The answer is the use of cutting-edge technologies such as big data, artificial intelligence, sensors, the Internet of Things (IoT), and computer cloud that enable 'excellent intelligence', 'the vastness of information' and 'interconnect between information and intelligence'. Is in. The technologies mentioned are in the spotlight as key technologies that will lead the Fourth Industrial Revolution. By constructing smart cities, countries and corporations combine urban public functions and infrastructure with these advanced technologies to minimize the social side effects and costs such as energy shortages and environmental pollution, while minimizing the power, convenience, and efficiency of information processing. We want to increase speed and achieve sustainable economic growth. It is hoped that the lives of citizens living in the city will naturally be improved in quality. Smart City is a smart city in that it is a space where work can be done quickly, conveniently, and efficiently.

However, the characteristics of such 'smart city' are in line with the main characteristics (modernity) of modern culture, and thus share some of the same problems. As explained above, the key to the current smart city is the use of advanced technologies related to the vastness of information, the excellence of intelligence, and the interconnection between information and intelligence. In this sense, 'intelligence', 'information', and 'connection' are key keywords that characterize smart cities. So, let's talk about the problems with keywords.

First, smart cities, which utilize a large amount of information, deepen the “privacy issue” raised by M. Foucault (1926–1984). Big data refers to the “voluminous amounts of data created in the digital environment," that is, large amounts of information. Advances in big data technology have enhanced artificial intelligence, which evolves through deep learning. That is, the more data, the greater the learning distance of artificial intelligence, and the superiority of artificial intelligence with increased learning amount increases in direct proportion. Therefore, for a smart city to fulfill its role as a literally smart city, the more data (information) that is the learning distance of artificial intelligence, the better. However, with the development of sensors and the Internet of Things (IoT), the collection of such data has become easier. According to the latest smart city policy, sensors and artificial intelligence and information and communication technology (ICT) are installed in almost every place where citizens living in the city are trying to collect information in real time. The problem, however, is that, in the pursuit of convenience, speed and efficiency in work processing, relying heavily on advanced science and technology, citizens' personal information can be easily exposed in various parts of the city and their lives can be controlled. On the other hand, Foucault criticized the nature of modern culture, disregarding the privacy and rights of individuals, while being buried only in pursuit of 'orderliness', 'safety' and 'efficiency'. This privacy issue, however, is likely to deepen in smart cities, which aim to collect and efficiently process personal information of citizens every moment. For example, if existing CCTVs were simply storing images of our faces and movements as images, in this new city, we could analyze every taste, habit, and lifestyle that we didn't even know about ourselves. You may be hit. Let's look at specific examples regarding privacy issues. China's CCTV and UNHCR refugee data.

CCTV in China: Shenzhen, which is the first "special economic zone" in China, is equipped with more than 200,000 CCTVs for preparation for Beijing Olympics and they are installed throughout the city and are disguised as lampposts. These CCTVs are for high-tech surveillance and censorship program in China, called "Golden Shield", and the system is supplied by IBM, Honeywell, and GE in North America. In 2006, China installed security cameras connected directly to police stations for "Safe Cities" surveillance project in all restaurants and internet cafes, and these are now related to "Golden Shield," watching Chinese citizens around the clock throughout networked CCTVs, monitored by digital voice-recognition technologies, and limited internet access throughout online controls known as the "Great Firewall." Citizens' movements will be tracked through national ID systems with computer chips, and photos taken with CCTV systems are uploaded to police databases and linked to the holder's personal data, and this links all this information together in a searchable database of personal information. This system helps prevent thieves such as stealing GPS from the car as well as sudden political outbreaks, such as by identifying demonstrators and rioters during the outbreaks and posting them online to help with the manhunt [4].

UNHCR refugee data: Since the digital tools such as social network services and online messaging became widely available, refugees were able to reply on digital networks to communicate with distant families and locate the resources they need, but the same tools that gave refugees freedom can easily be leverage for surveillance and control over refugees. In the recent context of the Syrian refugee crisis, the use of digital media and communication technologies got attention. With digital infrastructure such as Wi-Fi hotspots, cell phone charging stations and mobile apps, asylum seekers rely on digital devices in order to navigate the external and internal borders of the EU nations, making success of refugees in making it to safe spaces rely on not only physical but also digital infrastructure. But the very same technologies they use can be utilized by human traffickers who leverage digital media to operate efficiently. Also, governments took advantage of social media and digital biometrics to identify and control their borders, making digital platforms facilitating movement simultaneously sites of surveillance that instantiate government border policies of control[5].

Second, there is a problem that 'intelligence', which is important in smart cities, is defined in a very narrow sense. Intelligence emphasized in smart city policy is closely related to artificial intelligence technology. Artificial intelligence, which is being developed based on human intelligence (reasonable ability), is now solving the problem by calculating numerical values faster than humans or by efficiently classifying data such as text, voice, pictures, and images. Because of this, if artificial intelligence is to be exerted throughout the city in the future, it may be easier for citizens to do their jobs. However, intelligence emphasized in smart city policy cannot be said to represent all human intelligence. As C. Taylor (1931--) gives a hint, today's artificial intelligence is built from the very beginning, emulating only the "mechanical cognition, computation, and procedural reasoning" capabilities that are part of human intelligence[5]. This narrowly defined intelligence, no matter how excellent it is, will naturally improve the quality of life of citizens (in other words, lead to 'better life', 'higher life' and 'better life'.) Cannot be seen. As M. Horkheimer (1895-1973) criticized, the quality or value of human life is by no means an invaluable part of 'tool reason' of economic efficiency. Rather, the evaluation of quality of life and the orientation to quality of life can be carried out by intelligence different from mechanical intelligence — in Taylor's way, self-interpretation. However, recent smart city policies overlook this.

Third, smart cities that try to connect everything can give rise to 'totalitarianism'. In smart cities, all things and people are connected. As mentioned earlier, the smartness of a smart city (i.e. the prominence of artificial intelligence) is proportional to the amount of information. That's why scientists and policymakers are trying to bring together all information and intelligence through computer cloud technology. In other words, the information provided by citizens is stored and used in a single data center (central cloud) to facilitate information processing and to promote the city's smartness quickly. The link between information and intelligence is not always accompanied by the consent of individual citizens. This allows citizens to unknowingly expose their secret information. Even if this connection is done with agreement, there is a problem. Smart City, where everything is connected, poses the danger of "totalitarianism" that uniformizes the way of life of individual citizens. Because citizens living in smart cities will be provided with information in the same data center, they are likely to enjoy a similar way of life. In addition, there is the possibility of excluding life intentionally rejecting this connection or inevitably falling apart. That is to say, as modern culture has treated the non-mainstream life out of the mainstream, and has ruled out the whole, 'disconnected life' for any reason in a smart city—in a smart city, that is, relatively lack of information or intelligence, you will end up with a slow and inefficient life—you can be alienated or treated relatively low. But fear of such discrimination and inequality leads to forced and violent circumstances.

4. Conclusion

No one may like a city that is too smart, as criticized by sociologist R. Sennett[6]. Also, if someone really wants a smart city, someone else may want a different city. Again, with help from Taylor, we are intelligent beings who can evaluate our own lives. In the smart city, however, intelligence is a narrowly defined concept, so no matter how highly intelligent the city, it can be expected that the quality of life of citizens will increase in direct proportion. Rather, the qualitative fullness of life (the condition and condition of a good life) should be judged as an intelligence that can be defined in a broader sense, an intelligence that only humans can possess. And its appearance and conditions can be judged differently depending on how individual citizens exercise their intelligence. Therefore, even if we live in the same space, our ways of life are bound to be colorful. Nevertheless, today's smart cities (an urbanization policy emphasized by the term “smart city”) are overlooked. Currently, the meaning of smart in smart cities is being discussed focusing on human convenience, energy efficiency and utility. But what will be discussed importantly in smart cities is what the city means and for whom it should be smart. In other words, what is the purpose of smart technology?

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A6A3A01078538).

References

1. D. McFarlanea, S. Sarmab, J. L. Chirna, C. . Wonga, and K. Ashton, “Auto ID systems and intelligent manufacturing control,” Engineering Applications of Artificial Intelligence, vol. 16, no. 4, pp. 365–376, (2003)
2. ITU, Smart sustainable cities: An analysis of definitions, ITU-T Focus Group on Smart Sustainable Cities. (2014).
3. Klein, N., China’s all-seeing eye. Rolling Stone, 28 (2008).
4. Latonero, M., & Kift, P. On digital passages and borders: Refugees and the new infrastructure for movement and control. Social Media+ Society, 4(1), 2056305118764432 (2018).
5. C. Taylor, Sources of the Self: The Making of the Modern Identity, Cambridge, Mass.: Harvard University Press, pp. 72-75 (1989).
6. Richard Sennett, “No one likes a city that's too smart”(2012.12.4), The Guadian(search date: 2020.1.8) https://www.theguardian.com/commentisfree/2012/dec/04/smart-city-rio-songdo-masdar

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.019

An Ethical Black Box for Social Robots: a draft Open Standard

Alan F.T. Winfield[1*], Anouk van Maris[1], Pericle Salvini[2], and Marina Jirotka[2]

[1] Bristol Robotics Lab, University of the West of England, Bristol
[2] Department of Computer Science, University of Oxford

Abstract. This paper introduces a draft open standard for the robot equivalent of an aircraft flight data recorder, which we call an ethical black box. This is a device, or software module, capable of securely recording operational data (sensor, actuator and control decisions) for a social robot, in order to support the investigation of accidents or near-miss incidents. The open standard, presented as an annex to this paper, is offered as a first draft for discussion within the robot ethics community. Our intention is to publish further drafts following feedback, in the hope that the standard will become a useful reference for social robot designers, operators and robot accident/incident investigators.

Keywords: ethical black box, social robots, traceability, transparency, robot ethics, responsible robotics

1 Introduction

In [8] we argued the case that robots and autonomous systems should be equipped with the equivalent of an aircraft Flight Data Recorder to continuously record sensor and relevant internal status data. We call this an ethical black box (EBB). We argued that an ethical black box will play a key role in the processes of discovering why and how a robot caused an accident, and thus an essential part of establishing accountability and responsibility.

We propose that the EBB needs a standard specification. A standard specification has several benefits. First, a standard approach to EBB implementation in social robots will greatly benefit accident and incident (near miss) investigations [9]. Second, an EBB will provide social robot designers and operators with data on robot use that can support both debugging and functional improvements to the robot. Third, an EBB can be used to support robot 'explainability' functions to allow, for instance, the robot to answer 'Why did you just do that?' questions from its user. And fourth, a standard allows EBB implementations to be readily shared and adapted for different robots and, we hope, encourage manufacturers to develop and market general purpose robot EBBs.

This paper is structured as follows. Section 2 provides a brief recap of the history of data loggers, and associated standards, in aviation, critical infrastructure

* alan.winfield@brl.ac.uk

and road vehicles. In section 3 we give a brief high-level description of the EBB, and in section 4 we argue the case for a standardised EBB, and a draft open standard as a starting point. In section 5 we conclude the paper by outlining the draft open standard in Annex A.

2 A brief introduction to Data Loggers

The term 'black box' was first used informally in the late 1940s for navigational instruments, within the Royal Air Force. The term was then extended to cover any kind of apparatus within a sealed container. From the mid 1960s the colloquial use of the 'black box' has narrowed to refer to the Flight Data Recorder (FDR), now fitted as standard in aircraft.

Black box – or flight data recorders – were introduced in 1958, for larger aircraft, and since then have vastly expanded in scope in what flight data they record. Initially FDRs included time navigation data about the position of surfaces and the pilots' movement of controls; latterly sensor data on the internal and external environment as well as the functioning of components and systems are also recorded, alongside autopilot settings such as selected headings, speeds, altitudes and so on [2]. FDRs on modern aircraft record more than 1000 parameters.

The transfer of the black box concept into settings other than aviation is not new. Data loggers for critical infrastructure such as Supervisory Command and Data Acquisition (SCADA) systems are also standard practice [5, 6]. The largest deployment of black box technology outside aviation is within the automobile and road haulage industries for data logging [7, 10]. Data loggers for vehicles are generally known as Event Data Recorders (EDRs). Standards for EDRs include IEEE 1616 'Standard for Motor Vehicle Event Data Recorder (MVEDR)' first published in 2004 and revised in 2021 [1].

3 The Ethical Black Box

All robots collect sense data, and – on the basis of that sense data and some internal decision making process (AI) – send commands to actuators. This is of course a simplification of what in practice will be a complex set of connected systems and processes but, at an abstract level, all intelligent robots will have the three major subsystems shown in blue, in Fig. 1. A social robot is no different, except that it is designed to interact directly with humans.

We define the ethical black box (EBB) as a system for securely recording date- and time-stamped operational data from a social robot. We use the term *ethical black box* to emphasise our view that to deploy social robots without an EBB would be irresponsible.

The Ethical Black Box (EBB) and its data flows, shown in red in Fig. 1 will need to collect and store data from all three robot subsystems: sensor data, actuator demands and actual positions, and robot decisions — ideally with the reasons for those decisions. All of these data will need to be date and time

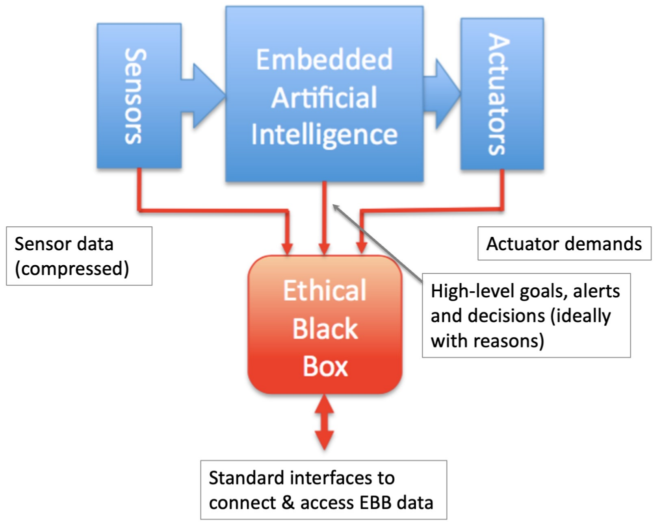


Figure 1. Robot sub-systems with an Ethical Black Box and key dataflows.

stamped. An important property of the EBB is that the data flows from the robot to the EBB are strictly one-way. It is important that the EBB is a passive subsystem, accepting data from the robot controller, and not affecting the robot's operation.

4 Why do we need a standardised EBB?

We contend that social robots should not only be fitted with an EBB, but that EBB should follow a standard specification. There are several benefits from a standard EBB:

1. A standard approach to EBB implementation in social robots will greatly benefit accident and incident (near-miss) investigations. In [9] we argue that social robots bring greater risks than industrial robots, and hence the likelihood of harms — including psychological, societal or environmental harms — is greater. Without the data provided by the EBB the investigation of accidents or near-miss incidents in order to discover what happened, why it happened and how to prevent it happening again is difficult, if not impossible.
2. An EBB will provide social robot designers and operators with data on robot use that can support both debugging and functional improvements to the robot. Thus it becomes a powerful diagnostic tool during research and development.

3. An EBB can be used to support robot explainability functions to allow, for instance, the robot to answer "Why did you just do that?" questions from its user [4].[3] And
4. a standard allows EBB implementations to be readily shared and adapted for different robots and, we hope, both encourage manufacturers to develop and market general purpose robot EBBs, and regulators to require EBBs.

Bruce Perens, creator of The Open Source Definition, outlines a number of criteria an open standard must satisfy, including:

1. "Availability: Open standards are available for all to read and implement.
2. Maximize End-User Choice: Open Standards create a fair, competitive market for implementations of the standard.
3. No Royalty: Open standards are free for all to implement, with no royalty or fee.
4. No Discrimination: Open standards and the organizations that administer them do not favor one implementor over another for any reason other than the technical standards compliance of a vendor's implementation.
5. Extension or Subset: Implementations of open standards may be extended, or offered in subset form."[4]

5 The Draft Open Standard

In Annex A we set out the first draft of an Open Standard for an EBB for social robots. Following the model of the Internet open standards this draft is a Request for Comments (RFC). Subsequent drafts will incorporate feedback. The standard is written following BS 0:2021 *A Standard for Standards* [3].

Annex A sets out the requirements for an EBB Software Module that could be integrated within a robot's controller, or as a stand alone software process connected, via a network connection, to the robot.

Annex A details the normative requirements for the data structures and formatting. These data structures fall into three categories: Meta Data, which stores information about the robot that the EBB is connected to, Data Data, which stores information on the number of records and dates and times of the oldest and most recent records in the EBB, and Robot Data, which stores operational data on the robot. The final category Robot Data will comprise most of the data stored in the EBB. All three structures are date- and time-stamped and have a common format. Annex A also provides examples of records for each of these three data structures together with an example of a complete set of EBB records, and how often particular robot data records might be written.

The overall aim of the standard is to provide a technical specification for an EBB for Social Robots. This specification, alongside an open-source library of model implementations, will provide a resource to enable developers to build an EBB into their robots.

[3] Noting that this would require the robot's control system to access it's own EBB - and thus violate the principle of the EBB's passivity.

[4] https://opensource.com/resources/what-are-open-standards

Acknowledgments

This work has been conducted within project RoboTIPS: Developing Responsible Robots for the Digital Economy supported by EPSRC grant ref EP/S005099/1.

References

1. IEEE standard for motor vehicle event data recorder (mvedr). *IEEE Std 1616-2021 (Revision of IEEE Std 1616-2004)*, pages 1–184, 2021.
2. D. R. Grossi. Aviation recorder overview, national transportation safety board [NTSB]. *J. Accid. Investig.*, 2(1):31–42, 2006.
3. British Standards Institute. BS 0 2021, a Standard for Standards — principles of standardization. *BS 0 2021*, pages 1–42, 2021.
4. Vincent J. Koeman, Louise A. Dennis, Matt Webster, Michael Fisher, and Koen Hindriks. The "why did you do that?" button: Answering why-questions for end users of robotic systems. In Louise A. Dennis, Rafael H. Bordini, and Yves Lespérance, editors, *Engineering Multi-Agent Systems*, pages 152–172, Cham, 2020. Springer International Publishing.
5. Thomas Morris and Kalyan Pavurapu. A retrofit network transaction data logger and intrusion detection system for transmission and distribution substations. In *2010 IEEE International Conference on Power and Energy*, pages 958–963, 2010.
6. Jeong Seok Oh. A practical study on data logger for gas industry. In James J. Park, Vincenzo Loia, Gangman Yi, and Yunsick Sung, editors, *Advances in Computer Science and Ubiquitous Computing*, pages 860–864, Singapore, 2018. Springer Singapore.
7. P. R. Thom and C. A. MacCarley. A spy under the hood: controlling risk and automotive EDR. *Risk Manag. Mag*, 55(2):22—26, 2008.
8. A. F. Winfield and M. Jirotka. The case for an ethical black box. In Y. Gao, S. Fallah, Y. Jin, and C. Lekakou, editors, *Towards Autonomous Robotic Systems (TAROS 2017) Lecture Notes in Computer Science Vol. 10454*, pages 262–273. Springer, Cham, 2017.
9. Alan F. T. Winfield, Katie Winkle, Helena Webb, Ulrik Lyngs, Marina Jirotka, and Carl Macrae. Robot Accident Investigation: A case study in responsible robotics. In Ana Cavalcanti, Brijesh Dongol, Rob Hierons, Jon Timmis, and Jim Woodcock, editors, *Software Engineering for Robotics*. Springer, Cham, 2021.
10. M. Worrell. Analysis of bruntingthorpe crash test data, impact. *J. Inst. Traffic Accid. Investigators*, 21(1):4–10, 2016.

Annex A: Draft Standard for an Ethical Black Box Software Module for Social Robots

Request For Comments, Draft 0.1

A.1 Scope

This draft open standard sets out normative technical requirements for a data logger for Social Robots, which we call an Ethical Black Box (EBB). This draft is a Request For Comments (RFC), inviting feedback and suggestions for improvements. Further drafts will incorporate revisions in response to this feedback.

The aim of the standard is to provide researchers, developers and operators with a technical specification for an EBB for Social Robots. This specification, alongside an open source library of model implementations, will provide a resource to enable developers to build an EBB into their robots.

The standard sets out the specification for a software module for an EBB, which may be implemented either as a software module alongside the robot's control system, or as a stand alone software process within a system connected (i.e. via a network connection) to the robot. This module may also be implemented within a hardware EBB physically connected to the robot, although the specification of the hardware is outside the scope of this draft.

A.2 Normative References

– BS 0:2021, *A standard for standards — Principles of standardization*
– BS ISO 8373:2021, *Robotics — Vocabulary*

A.3 Terms and Definitions

For the purposes of this document, the following terms and definitions apply.

Ethical Black Box (EBB): A system for securely recording date- and time-stamped operational data from a social robot.

Social Robot: An intelligent service robot designed to interact with humans. For definitions of intelligent service robot refer to BS ISO 8373:2021.

A.4 EBB Normative Requirements for Data

This section sets out requirements for the data that should be captured by the EBB.

EBB data organisation The data stored in the EBB shall comprise three types of record: meta data, data data and robot data. The EBB shall contain one meta data (MD) record only, one data data (DD) record only, and a number of robot data (RD) records. The maximum number of RD records shall be fixed for each EBB according to the limits of its storage capacity. The general organisation of EBB data is shown in Table 1.

The EBB RD logs shall be written in order starting from RD record 1, and proceeding to RD record n. After record n has been written the EBB shall write the next RD record overwriting RD record 1, and write subsequent RDs in record 2, and so on. In this way the EBB always stores the most recent set of RD records, up to its maximum capacity of n records.

Meta Data, Data data and Robot Data records have the same overall structure, as shown in 2.

Each EBB record consists of a 2 character record labels 'MD', 'DD' or 'RD' followed by a variable number of fields.

EBB records
Meta data (MD) record
Data data (DD) record
Robot data (RD) record 1
Robot data (RD) record 2
...
Robot data (RD) record n

Table 1. General organisation of EBB data

Record	length	format
Record type 'MD', 'DD' or 'RD'	2 chars	ASCII text
Number of fields and chars in record	12 chars	ASCII 000:00000000
Date record written	10 chars	ASCII yyyy:mm:dd
Time record written	12 chars	ASCII hh:mm:ss:ms
Data record 1	variable	see below
...		
Data record m	variable	see below
Checksum	4 chars	ASCII

Table 2. Common structure of Meta Data, Data Data and Robot Data Records

Each field in a record consists of a 4 character label, followed by data elements defined according to the label. EBB fields are defined below for each of the 3 types of record. Note that several fields are common to all 3 record types.

The EBB Meta Data Record The Meta Data Record shall store information about the robot that the EBB is fitted to (name, version or model no, and serial no), the robot's developer/manufacturer and operator, the contact details of the person responsible for the robot, and information on the EBB itself.

Table 3 defines each field in the MD record, and includes both required and optional fields. Here a string is defined as a variable length ASCII sequence terminated by the null character ASCII \0.

Notes on the MD fields:

1. The record size *recS*, has a 12 char data element formatted as 3 numeric chars and 8 numeric chars separated by a colon, i.e. 000:00000000. It follows that the maximum permissible number of fields in a record is 999, and the maximum number of characters is 99,999,999. Note also that the size and character count must include the recS field.
2. The EBB date field *ebbD* is 12 ASCII characters with colon separated year, month, and day, i.e. yyyy:mm:dd.
3. The EBB time field *ebbT* is 10 ASCII characters with colon separated hour, minute, second and millisecond, i.e. hh:mm:ss:ms.
4. The record checksum *chkS* shall be computed using a 64-bit non-cryptographic hash function, to be determined.

An example of a complete Meta Data Record is shown in 4.

label	data	length	requirement
recS	record size, field and chars, including recS field	12 chars	required
ebbD	EBB date record written	10 chars	required
ebbT	EBB time record written	12 chars	required
botN	robot name	string	required
botV	robot version no	string	optional
botS	robot serial no	string	optional
botM	robot manufacturer	string	required
opeR	robot operator	string	optional
resP	name and contact details of responsible person	string	required
ebbN	EBB name and version no	string	required
chkS	checksum for complete record	8 hexadecimal chars	required

Table 3. Meta Data Fields

field	comment
MD	record label
recS 010:00000000	number of fields:chars in record
ebbD 2022:04:20	date 20 April 2022
ebbT 16:40:20:000	time 16:40 and 20.000 seconds
botN NAO\0	NAO robot
botV 4\0	v4
botM Aldebaran\0	Manufacturer
opeR Bristol Robotics Lab\0	Operator
resP A Winfield +44 117 328 6913\0	person responsible
ebbN PyEBB v1.2\0	this EBB
chkS AF5679FC	checksum for this record

Table 4. An example Meta Data Record

The EBB Data Data Record The Data Data Record shall store information about the robot data records stored in the EBB.

Table 3 defines each field in the DD record, and includes both required and optional fields.

Notes on the DD fields:

1. For notes on *recS* *ebD1*, *ebT1*, *ebDM*, *ebTM* and *chkS* see notes on Table 3 above.
2. Field *ebbX* is an offset, in number of bytes, from the start of the EBB storage media to the next writable position for an RD record. Note that this will need to be reset back to RD 1 once the storage media is full.
3. Field *sysX* ‘system exclusive’ is manufacturer/operator definable. The data is formatted as 2 chars and a string separated by a colon, i.e. 00:string. The 2 characters are to allow the manufacturer to define up to 99 sysX fields, and the string allows for variable length data.
4. Given that all fields in the DD record are required and have a fixed length the *recS* field will have the default value 010:00000130, as shown in the example

label	data	length	requirement
recS	record size, field and chars	12 chars	required
ebbN	total number of EBB Data Records stored in EBB	10 chars	required
ebbX	index to the start of next writable RD record	16 chars	required
ebD1	date of oldest RD record written	10 chars	required
ebT1	time of oldest RD record written	12 chars	required
ebDM	date of most recent RD record written	10 chars	required
ebTM	time of most recent RD record written	12 chars	required
sysX	manufacturer definable field	variable	optional
chkS	Checksum for complete record	8 hexadecimal chars	required

Table 5. Data Data Fields

DD record in Table 6. Only if the record includes sysX field(s) will chkS have a different value.

An example of a complete Data Data Record is shown in 6.

field	comment
DD	record label
recS 010:000000130	number of fields:chars in record
ebbD 2022:04:20	date 20 April 2022
ebbT 16:40:20:000	time 16:40 and 20.000 seconds
ebbN 0000000400	400 records in EBB
ebbX 00000000001545060	offset to next RD position in storage media
ebD1 2022:03:01	first RD date 1 March 2022
ebT1 08:00:30:000	first RD time 08:00 and 30.000 seconds
ebDM 2022:05:01	Most recent RD date 1 May 2022
ebTM 18:59:30:100	Most recent RD time 18:59 and 30.100 seconds
chkS FF5678AC	checksum for this record

Table 6. An example Data Data Record

EBB Robot Data Records The Robot Data Records store operational data from the robot

Table 7 defines each field in the RD record, and includes both required and optional fields.

Notes on the RD fields:

1. For notes on *recS*, and *chkS* see notes on Table 3 above. For notes on *sysX* see notes on Table 5.
2. Field *botT* is needed in case the robot's clock shows a different time to the EBB's clock. The format of *botT* is the same as *ebbT*: 10 ASCII chars, for hours, minutes, seconds and milliseconds 00:00:00:000.

label	data	length	requirement
botT	robot time	10 chars	required
actD	actuator no and demand value	12 chars 000:±0000.00	optional
actV	actuator no and actual value	12 chars 000:±0000.00	optional
batL	battery level	3 chars	optional
tchS	touch sensor no and value	6 chars 00:000	optional
irSe	infra red sensor no and value	6 chars 00:000	optional
lfSe	line following sensor no and value	6 chars 00:000	optional
gyrV	gyro no and value	20 chars 00:±0000:±0000:±0000	optional
accV	accelerometer no and value	20 chars 00:±0000:±0000:±0000	optional
tmpV	temperature sensor no and value	8 chars 00:±0000	optional
micI	microphone no and input	variable, 2 chars:8 chars:wav hex	optional
camF	camera no and frame grab	variable, 2 chars:8 chars:jpg hex	optional
txtC	text input command	variable, string	optional
txtR	text reply	variable, string	optional
decC	robot decision code and reason	variable, 4 chars 0000:string	optional
wifi	WiFi status and signal strength	4 chars 0:00	optional
sysX	manufacturer definable field	variable, 2 chars 00:string	optional
chkS	checksum for complete record	8 hexadecimal chars	required

Table 7. Robot Data Fields

3. Fields *actD* and *actV* each contain values for the actuator number and the actuator demand, or actual positions respectively. The data is formatted 000:±0000.00 thus allowing for a maximum of 999 actuators, and positive or negative values of up to ±9999.99.
4. Fields *tchS, irSe, lfSe* each contain values for sensor number and the sensor value. The data is formatted 00:000 thus allowing for up to 99 sensors, and sensor values between 0...999.
5. Field *micI* allows for storage of a wav audio clip captured by the robot's microphone(s). The data has three colon separated elements, 2 chars for the microphone number, 8 chars for the length of the wav clip, and the hexadecimal representation of the wav binary. This allows for up to 99 microphones and wav clips of up to 99,999,999 bytes.
6. Field *camF* allows for storage of a jpg still frame grabbed by the robot's camera(s). The data has three colon separated elements, 2 chars for the camera number, 8 chars for the length of the jpg clip, and the hexadecimal representation of the jpg binary. This allows for up to 99 cameras and jpg clips of up to 99,999,999 bytes.
7. Fields *txtC* and *txtR* allow for storage of text input commands to the robot, and text responses from the robot, respectively. The input command might be typed by the robot's user, or spoken and then converted from speech to text by the robot's speech recognition system. The output text might be displayed visually or spoken by the robot's speech synthesis system. The data elements for both *txtC* and *txtR* are null terminated strings.

8. Field *decC* allows for storage of the robot's internal decision of its next action (i.e. turn left, turn right, stop, speak an alert, etc), together (optionally) with a reason for that decision. The data is formatting as a 4 char decision code, followed by a string. This allows for up to 9999 decisions to be logged. The determination of which numeric value to use for each robot decision is outside the scope of this standard, and left to the robot's manufacturer or operator. If there is no reason the string shall be stored as an empty null string terminated by ASCII \0.
9. Field *wifi* allows for storage of the robot's WiFi connection status and signal strength. The colon separated data is formatted as 1 character connection status, connected (1) or not connected (0), and 2 characters for signal strength, from 00..99.

An example of a complete Robot Data Record is shown in Table 6, for a simple differential drive wheeled robot, with 8 IR sensors.

field	comment
RD	record label
recS 017:00000nnn	number of fields:chars in record
ebbD 2022:04:20	RD date
ebbT 16:40:20:000	Rd time
batL 255	battery level
actV 001:-175.54	left wheel angle
actV 002:102.09	right wheel angle
irSe 001:0.05	IR sensor 1
irSe 002:0.05	IR sensor 2
irSe 003:0.05	IR sensor 3
irSe 004:0.05	IR sensor 4
irSe 005:0.05	IR sensor 5
irSe 006:0.23	IR sensor 6
irSe 007:0.15	IR sensor 7
irSe 008:0.05	IR sensor 8
decC 0020:obstacle detected\0	turning left to avoid obstacle on right
wifi 1:255	WiFi connected, good signal
chkS CFA3569A	checksum for this record

Table 8. An example Robot Data Record

A.4.2 EBB Timing Not all EBB records need to be written with the same frequency. In general those RD records which capture actuator movements and short range sensor inputs will need to be written with the highest frequency; a default setting for these might be once every 2 seconds. RD records that capture camera images might be written with a lower frequency, especially if the robot moves slowly, say once every 10 seconds, or less. A third group of RD records are those that capture aperiodic events, such as a user's commands and the robot's response (if there is one).

Table 9 illustrates the EBB timing, with three kinds of event logged: high frequency motor and sensor values captured once every 2 seconds (in RD 1, 2, 4, 6 etc), lower frequency camera frame grabs once every 10 seconds (in RD 3, RD 10 and RD 17), and the sporadic events of user commands to override the robot's autonomous operation, in RD 5 and RD 13. For clarity only the *ebbT* time fields are shown in full. Abbreviated *camF* and *txtC* fields are also shown in RD 3, 5, 10, 13 and 17.

label	fields
MD	recS... ebbD... ebbT 08:40:20:000 BotN ePuck BotM... Resp... chkS...
DD	recS... ebbD... ebbT 08:40:20:000 ebbD1... ebbT1... ebbN... ebDM... ebbTM... chkS...
RD 1	recS... ebbD... ebbT 08:40:22:000 botT... actV... batL... irSE... chkS...
RD 2	recS... ebbD... ebbT 08:40:24:000 botT... actV... batL... irSE... chkS...
RD 3	recS... ebbD... ebbT 08:40:25:000 botT... camF 01:00307200:... chkS...
RD 4	recS... ebbD... ebbT 08:40:26:000 botT... actV... batL... irSE... chkS...
RD 5	recS... ebbD... ebbT 08:40:27:100 botT... txtC Halt\0 chkS...
RD 6	recS... ebbD... ebbT 08:40:28:000 botT... actV... batL... irSE... chkS...
RD 7	recS... ebbD... ebbT 08:40:30:000 botT... actV... batL... irSE... chkS...
RD 8	recS... ebbD... ebbT 08:40:32:000 botT... actV... batL... irSE... chkS...
RD 9	recS... ebbD... ebbT 08:40:34:000 botT... actV... batL... irSE... chkS...
RD 10	recS... ebbD... ebbT 08:40:35:000 botT... camF 01:00307200:... chkS...
RD 11	recS... ebbD... ebbT 08:40:36:000 botT... actV... batL... irSE... chkS...
RD 12	recS... ebbD... ebbT 08:40:38:000 botT... actV... batL... irSE... chkS...
RD 13	recS... ebbD... ebbT 08:40:27:100 botT... txtC Run\0 chkS...
RD 14	recS... ebbD... ebbT 08:40:40:000 botT... actV... batL... irSE... chkS...
RD 15	recS... ebbD... ebbT 08:40:42:000 botT... actV... batL... irSE... chkS...
RD 16	recS... ebbD... ebbT 08:40:44:000 botT... actV... batL... irSE... chkS...
RD 17	recS... ebbD... ebbT 08:40:45:000 botT... camF 01:00307200:... chkS...

Table 9. Example EBB illustrating variable frequency of RD sampling

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.020

THE SHAPE OF ROBOT: APPEARANCE OF SOCIAL ROBOTS BASED ON A CASE STUDY

TIM CHEONGHO LEE

Sangmyung University
Seoul, South Korea
E-mail: humantad@hanmail.net
www.smu.ac.kr

The advent of social robots is expected to enhance human life in many respects. Social robots may satisfy basic human needs and enable a more convenient lifestyle. One of the basic human needs is the desire for social life which inevitably requires social robots to equip with communication skills that enables close relationships with humans. However, in addition to communication skills, the appearance of social robots may play a pivotal role in shaping close relationships between humans and robots.

This may be especially important when social robots are utilized for the education of elementary or secondary school students. If social robots have a "sufficiently human shape," students may experience social robots with familiarity, identification, and empathy.[1] This is because social robots in human shape could be a source of curiosity and/or fantasy, facilitate students' understanding due to the social robot's ability to imitate, and support conversations (verbal and physical communication) in a non-judgmental and realistic perceptual way.

All of these may not be the case if social robots are "appropriately human." The appropriately human robots may not fulfill the educational purpose due to the psychological distance between humans and robots. Students may feel robots like a machine that does not deserve sympathy.[2] However, student may feel less stressful with the appropriately human robots because students can be relieved from stress because of their perception of robots to be an inhuman machine that cannot judge itself and thus feel less intimidating and less fear of making mistakes.

Ethical issues may arise when using social robots in education. Students may experience excessive emotional attach to robots.[3] For this, they are required to be remined of the inhumane characteristics of social robots. Students also could harass against humanoid social robots. There could be privacy issues as well including information related to an individual's education status. Social robots are not free from the problem of bias when they make recommendations after searching for data. Still, social robots may not be conveniently accessible to some people and gaps may occur due to economic status, generation, race, cultural difference, etc.

References

1. T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, F. Tanaka (2018). "Social robots for education: A review," *Science Robotics*, *3* (21).2018; Papadopoulos et al., 2020.
2. Irena Papadopoulos, Runa Lazzarino, Syed Miah, Tim Weaver, Bernadette Thomas, Christina Koulouglioti (2020). "A systematic review of the literature regarding socially assistive robots in pre-tertiary education," *Computers & Education*.
3. Papadopoulos et al. (2020).

SECTION–5

HUMAN-ROBOT INTERACTION

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.005

CASE-BASED ROBOTIC ARCHITECTURE WITH MULTIPLE UNDERLYING ETHICAL FRAMEWORKS FOR HUMAN-ROBOT INTERACTION

SHENGKANG CHEN, RONALD C. ARKIN, JASON BORENSTEIN
Georgia Institute of Technology, Atlanta, GA, USA
schen754@gatech.edu arkin@cc.gatech.edu, borenstein@gatech.edu

ALAN R. WAGNER, and ARTHUR MELO CRUZ
Pennsylvania State University, University Park, PA, USA
alan.r.wagner@psu.,edu, amc6630@psu.edu

As robots are becoming more intelligent and more commonly used, it is critical for robots to behave ethically in human-robot interactions. However, there is a lack of agreement on a correct moral theory to guide human behavior, let alone robots. This paper introduces a robotic architecture that leverages cases drawn from different ethical frameworks to guide the ethical decision-making process and select the appropriate robotic action based on the specific situation. We also present an architecture implementation design used on a pill sorting task for older adults, where the robot needs to decide if it is appropriate to provide false encouragement so that the adults continue to be engaged in the training task.

1. Introduction

Making ethical decisions is challenging but it is something the people have to do regularly in their daily lives. Robots may need to have the ability to make similar decisions within the context of human-robot interactions. In real-world situations, people follow different ethical rules and change their ethical decisions according to the situations. Since there is a lack of agreement on a unified ethical framework for human-human interactions, it is likely impractical to develop a unified single ethical framework appropriate for use in human-robot interactions. Moreover, if factors such as moral emotions affect a human's ethical decision-making process, robots may need to be able to make ethical decisions depending on the current emotional context to develop more meaningful human-robot relationships. In this paper, we describe a flexible robotic architecture with cases derived from different ethical frameworks, which potentially allows a robot to produce morally acceptable actions based on the selected ethical framework and the current situation.

2. Background

As robots are deployed in various fields and become more autonomous, human-robot interactions (HRI) are becoming more common. Researchers are noticing the possible ethical issues related to HRI and the need to develop ethical robots [1–4]. Various robotic architectures have been proposed for ethical behaviors [5–7]. In [5], the authors developed a robotic architecture to produce ethical behaviors based on predefined ethical rules and applied it to caregiving scenarios [8]. However, robots using this approach may be limited to well-characterized environments and well-defined rules derived by ethics experts. Alternatively, Abel, MacGlashan, and Littman [6] leverage reinforcement learning to allow robots to learn ethical behaviors, but they found that robots may behave inappropriately in unseen environments. Vanderelst et al. propose an architecture that uses forward simulation with a human model to evaluate possible robotic behaviors in order to find an appropriate one [7].

However, this model requires accurate human models which may not be readily accessible in many real-world scenarios.

Robotic deception has been an important topic in HRI [1]. Some researchers are concerned about the possible harmful impacts of deception in social robots [9–11]. One of the concerns is that users might overtrust a robot's capabilities and allow the robot to make unqualified decisions [9]. Moreover, Wilson et al. are concerned that robot deception may damage human-robot trust and can even lead to manipulation, especially for aging high-risk populations [11].

Other researchers believe robotic deception can be beneficial to human users [12, 13]. The authors in [12] found that deceptive behaviors of robots allow human users to be more engaged in game-play scenarios. In [13], the authors argue it is ethical for a robot to deceive if it benefits the overall human-robot relationship. To study people's opinions toward robotic deception, researchers distributed a questionnaire. They concluded that although deceptive behaviors decrease human trust in robots, the majority of the participants consider deception acceptable if these behaviors are beneficial to them [14]. However, this study was only limited to low-risk populations.

3. Architecture Design

This paper describes ongoing research [15] with an updated architecture for ethical robotic behavior. The goal of this architecture is to enable a robot to use various ethical frameworks for more robust ethical decision-making in HRI. It aims to produce morally acceptable behaviors in terms of experiences and outcomes for human users in complex real-world environments.

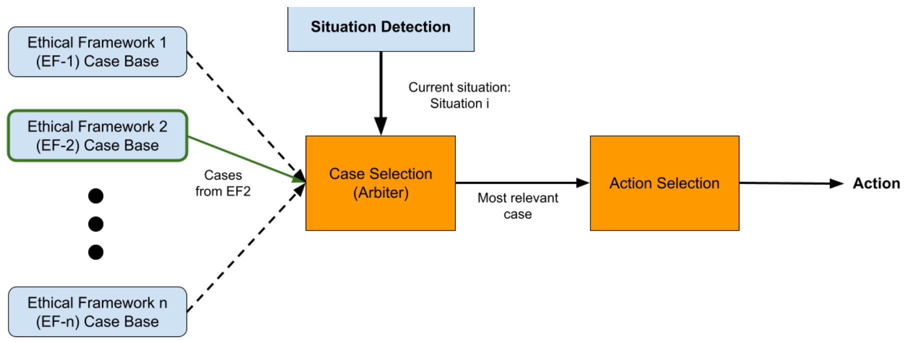


Figure 1. An overview of the architecture for ethical robotic behaviors. Given cases for a selected ethical framework (ethical framework 2 in this figure), the case selection module (arbiter) will select the most relevant case based on the information about the current situation. Then, the action selection module will choose the most appropriate action guided by the most relevant case.

The architecture utilizes the case-based reasoning (CBR) approach, a simple but effective methodology for artificial intelligent agent decision making [16]. In CBR, the robot uses information about the current situation based on decisions made in a similar previous situation from its case base. For case selection, the architecture (Figure 1) contains multiple cases for each ethical framework. Each set of cases contains cases drawn from surveys of people (either laypersons or ethics experts) on their opinions for different situations involving deception. The intent is to ensure that the robot's actions will be consistent with human moral decisions since these actions are guided by cases of human opinions. Each case in the case base is indexed by high-level features about the situation so the architecture can utilize the information about the current situation to find the most relevant case.

When deciding an appropriate action to perform, the robot's case selection module will find the most relevant case for a chosen ethical framework (derived in advance) using similarity measures between the current situation and the case indices (e.g., the Euclidean distances between the feature vectors). Provided with the most relevant case, the action selection module will then output an appropriate action for the robot to execute.

4. Architecture Implementation

This section presents an implementation of the robotic architecture for a specific human-robot interaction scenario: pill sorting with an older adult. Taking medications is part of the daily routine for many older adults, and pill sorting accuracy can be crucially important. However, pill sorting can be challenging for older adults with memory issues and training of the task can lead to frustration. During training, a robot observes and provides feedback about the older adult's performance on the task. In this pill soring scenario, we want to study whether it is moral to deceive an adult in a pill sorting task to keep them engaged with the task. Using the robotic architecture, the robot needs to decide whether to provide false encouragement (deception) or an accurate assessment (truth).

For the ethical framework cases, we considered ethical choices from both regular adults ("folk morality") and formal ethical frameworks: Utilitarianism [17], Kantian Ethics [18], Social Justice Theory [19], Ethics of Care [20], and Virtue Ethics [21]. To create the case base, we conducted two separate survey studies. For folk morality, 100 survey responses were collected through Amazon's Mechanical Turk service in January 2020. For the five formal ethical frameworks, 30 ethics experts were invited to answer the survey and 22 valid responses were received in February 2020. The survey data were then analyzed and used to create cases to populate each corresponding ethical framework. Each case contains the action probabilities derived from the survey data and is indexed by two binary variables (task performance and subject emotional state).

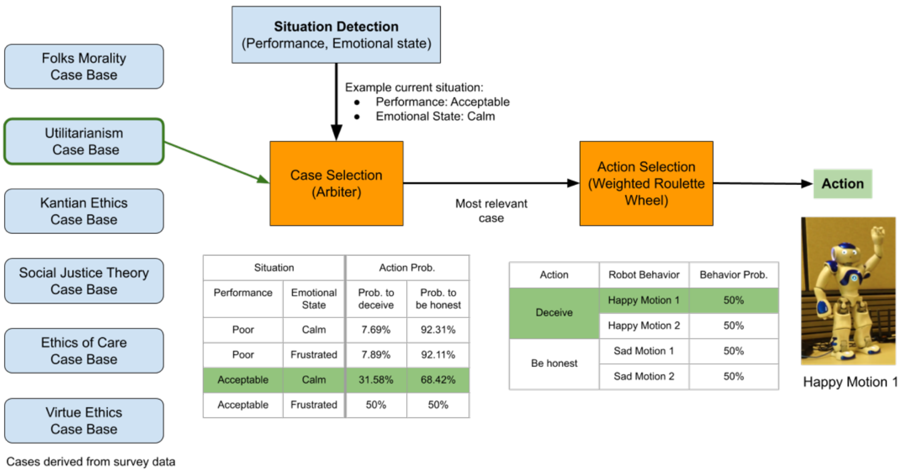


Figure 2. The initial architectural implementation for human-robot interaction in the pill sorting scenario with an NAO robot. The cases are derived from survey data. In this example, the case selection module uses Utilitarianism cases. For example, consider the older adult just made a mistake in a pill sorting task. The situation detection module provides the case selection module the information about the current situation (task performance and emotional state): here, the human user has an acceptable task performance and remains calm. Then, the case selection module finds the most relevant case (highlighted in green) and sends it to the action selection module. In this case, the action selection module chooses a deceptive action using a weighted roulette wheel selection. Happy Motion 1 is selected for the NAO robot to perform in the presence of the user providing false feedback on their result. In this case, the NAO robot deceives the older adult by providing false encouragement.

In the pill sorting task, the robot relies on its ethical architecture (Figure 2) to produce morally acceptable actions based on the chosen ethical framework and current situation. Using this implementation, the case selection module selects the most relevant case for a chosen ethical framework based on the current situation (user task performance and user emotional state). To avoid repetitive robotic behaviors, the robot will use behavior probabilities to generate different behaviors/gestures corresponding to an action (e.g., deceive). Using the action probabilities (derived from survey data) and behavior probabilities (defined by researchers), the NAO robot outputs an action by performing a gesture (e.g., happy motion 1) using the roulette wheel selection method [22] to provide feedback to the older adult on pill sorting results.

Currently, we only consider two binary variables to describe the situation: user performance (acceptable vs. poor) and emotional state (calm vs. frustrated). However, this architecture can easily be extended to more descriptive variables (e.g., scalar variables or categorical variables) with an updated case base. Moreover, more gestures for the NAO robots can be added to make the human-robot interaction process more engaging.

5. Discussion

In a real-world or ethically complex situation, it may not be appropriate to ask the robot to follow a set of fixed ethical rules regardless of the situation. Humans make different ethical decisions affected by differing situations. Thus, it is crucial for robots to be sensitive to the current context if they are going to be able to perform appropriate ethical actions during human-robot interactions. Consequently, this architecture allows the robot to produce appropriate actions based on a selected ethical framework and the current circumstances within which the user is situated. Moreover, the cases of the architecture can be expanded continuously during human-robot interactions by learning and adding new cases, a hallmark of case-based reasoning, which makes the robots more adaptive. This is crucial for building sustainable human-robot relationships.

A novel extension is to incorporate moral emotions into the architecture. Moral emotions [23] (e.g., guilt, empathy and anger) has been shown to play an important role in human ethical decision-making process [24, 25]. As a result, robots also need to take into consideration moral emotions in order to effectively support the human decision-making process.

Currently, we are implementing the architecture and plan to test it on physical robots. We want to conduct a series of HRI studies to evaluate the robotic architecture for two scenarios: pill sorting with an older adult and game playing with a child. We want to investigate whether the generated robotic actions using the architecture are morally acceptable to people under various ethical frameworks in different situations involving human participants, ideally by having an individual interact with a robot, but also through the use of focus groups.

6. Conclusion

In this paper, we present a flexible robotic architecture using a case-based reasoning approach for the generation of ethical behaviors consistent with either folk morality or decisions recommended by ethics experts for use in human-robot interaction. Moreover, we describe an architectural implementation for a specific human-robot interaction scenario: pill sorting with an older adult. In this scenario, the robot needs to decide whether to deceive the older adult by providing false encouragement to allow the older adult to continue the task or instead be honest by providing actual assessment results with the potential consequence of the user discontinuing the training due to frustration. We used the results of survey studies from both regular adults and ethics experts to generate various ethical framework cases that guide the decision-making process to produce appropriate actions relevant to the current situation. This architecture aims to become a tool for researchers to investigate further how to enable robots to interact with humans ethically.

Acknowledgments

This research project is supported by the National Science Foundation as part of the Smart and Autonomous Systems program under Grants No. 1849068 and 1848974. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. R. Wullenkord and F. Eyssel, *Current Robotics Reports 2020 1:3* 1, 85 (2020).
2. M. Anderson and S. L. Anderson, *AI Magazine* 28, 15 (2007).
3. M. Scheutz and B. F. Malle, *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, ETHICS 2014* (2014).doi:10.1109/ETHICS.2014.6893457
4. A. Leveringhaus, *European View* 17, 37 (2018).
5. R. C. Arkin and P. Ulam, An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions, (2009), pp. 381–387.doi:10.1109/CIRA.2009.5423177
6. D. Abel, J. MacGlashan, and M. L. Littman, *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence* (2016).
7. D. Vanderelst and A. Winfield, *Cognitive Systems Research* 48, 56 (2018).
8. J. Shim, R. Arkin, and M. Pettinatti, *Proceedings - IEEE International Conference on Robotics and Automation* 2936 (2017).doi:10.1109/ICRA.2017.7989340
9. A. Sharkey and N. Sharkey, *Ethics and Information Technology 2020 23:3* 23, 309 (2020).
10. J. Danaher, *Ethics and Information Technology 2020 22:2* 22, 117 (2020).
11. J. R. Wilson, M. Scheutz, and G. Briggs, 377 (2016).doi:10.1007/978-3-319-31413-6_18
12. E. Short, J. Hart, M. Vu, and B. Scassellati, No fair!! An interaction with a cheating robot, (2010), pp. 219–226.doi:10.1109/HRI.2010.5453193
13. J. Shim and R. C. Arkin, *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013* 2328 (2013).doi:10.1109/SMC.2013.398
14. K. Rogers and A. Howard, *Proceedings of IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO* 2021-July, 200 (2021).
15. R. C. Arkin, J. Borenstein, and A. R. Wagner, Competing ethical frameworks mediated by moral emotions in HRI: Motivations, background, and approach, (2019).doi:10.13180/ICRES.2019.29-30.07.001
16. J. L. Kolodner, *Artificial Intelligence Review 1992 6:1* 6, 3 (1992).
17. J. Driver, The History of Utilitarianism, in *The Stanford Encyclopedia of Philosophy*, Edited by E. N. Zalta, Metaphysics Research Lab, Stanford University (2014).
18. R. Johnson and A. Cureton, Kant's Moral Philosophy, in *The Stanford Encyclopedia of Philosophy*, Edited by E. N. Zalta, Metaphysics Research Lab, Stanford University (2022).
19. J. Rawls, *A theory of justice*, Cambridge, Massachusetts, The Belknap Press of Harvard University Press, (1971).
20. Virginia. Held, *The ethics of care personal, political, and global*, Oxford University Press (2005).
21. R. Hursthouse and G. Pettigrove, Virtue Ethics, in *The Stanford Encyclopedia of Philosophy*, Edited by E. N. Zalta, Metaphysics Research Lab, Stanford University (2018).

22. D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Professional (1989).
23. J. Haidt, THE MORAL EMOTIONS, in *Handbook of affective sciences*, Oxford University Press (2003).
24. J. P. Tangney, J. Stuewig, and D. J. Mashek, *Annual Review of Psychology* 58, 345 (2007).
25. C. D. Cameron, K. A. Lindquist, and K. Gray, *Personality and Social Psychology Review* 19, 371 (2015).

ICRES 2022: 7th International Conference on Robot Ethics
and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.014

ETHICAL SELF-DRIVING VS DATA ETHICS: THE NORMS OF COORDINATION AND THEIR CONSEQUENCES

KUMIN MOON

Humanities Research Institute, Chung-ang University, #828, 310 Hall, 84 Heukseok-ro, Dongjak-gu, Seoul, 06974, Korea

E-mail: dkxnaks@cau.ac.kr

We are living in the age of self-driving. Level 3 self-driving cars are already released all over the world, and now many companies are struggling to develop level 4. In level 3 and 4 self-driving cars, a special interaction takes place between level self-driving cars and their drivers. Drivers and automated driving systems(ADS) give and take control of their vehicles. The process of driver's taking control back from ADS is called takeover. Engineers and researchers agree that takeover is one of the most critical issues in developing and commercializing safer level 3 and 4 cars. The topic of this paper is the moral consequences of takeover in self-driving cars. First, technical and normative issues of takeover are presented. Since takeover may seriously affect the safety of both drivers and others, takeover itself raises normative concerns. To protect drivers' and others' lives and properties, what should ADS and drivers do? To answer this concern, I will address this concern by summarizing what might be called 'the norms of coordination'. They consist of two norms, one for ADS and another for drivers. Second, it will be argued that fulfilling one of those norms inevitably leads to collecting and storing data from drivers and their vehicles. This kind of data collecting makes takeover a matter of data ethics because it may be considered as micro-surveillance on drivers and their driving activities. Here, making self-driving ethical creates tension with data ethics: for ethical takeover, data collection on drivers and their driving activities is technically inevitable. But such data collection increases the risk of micro-surveillance. As a conclusion, I argue data collection from self-driving cars should be allowed. Such data will accelerate the progress toward more efficient and safer takeover technology. However, the collected data must be used only for research and development and protected from any hacking or unethical use. To this end, data ethics tailor-made for self-driving should be developed. Based on data ethics for self-driving, data privacy laws concerning self-driving vehicles can be legislated.

1. Ethics of self-driving: takeover and the norms of coordination

About a year ago, Honda Motors released Legend, their own model of level 3 self-driving car. Many of those who have been eagerly waiting for real, genuine self-driving cars paid attention to Legend, since it was the first level 3 self-driving car approved by SAE (Society of Automotive Engineers). Though some of the big tech companies like Tesla, Google, or Waymo were advertising themselves as leaders of the self-driving industry, their models were not qualified as level 3 by the SAE standard until Legend was released.[1] Of course, SAE is just one association among others, but the level system of self-driving they established is most widely accepted both in academia and the field. After the release of Legend, nearly every week we can hear news about level 4 or even level 5 self-driving cars. In technology, nobody can be certain about anything. But one thing is clear. The age of self-driving has already arrived. We are living in it.

Instead of unguarded optimism or persistent skepticism about self-driving, I will modestly focus on real-world cases at hand: level 3 and 4 self-driving cars. Technically, level 0, 1, and 2 are not qualified as self-driving cars. Level 0 cars are fully controlled by a driver to perform all driving tasks. There is no automation. Level 1 vehicles merely assist the driver. They can control only the speed or steering but cannot control both at the same time. Level 2 cars differ from

[1] For the latest version, see [1].

level 1 in that they can perform the two primary lateral and longitudinal controls. Level 3 cars can take all the driving tasks on their own under limited conditions. In automated driving mode, the driver does not have to take the wheel or step on pedals. All the driver has to do is to remain alert and be ready to take back control when the vehicle signals for it. Level 3 and 4 differ in only one aspect. When things go wrong or there is a system failure, level 4 vehicles can intervene, but level 3 vehicles cannot. Due to the lack of legislation and infrastructure, in the real world, their automated driving mode can be used under very limited conditions. Level 5 cars are fully automated. They do not need to have even steering wheels and pedals. They require zero human attention and will be able to go anywhere and do anything as human drivers can do. Level 0~2 cars are not self-driving. There is still a long way to go for level 5 to be realistic. For now, self-driving cars are qualified as level 3 or level 4 at best. In this article, 'self-driving cars' basically means level 3 and 4 self-driving cars.

The topic of this article is the moral consequences of special interactions between human drivers and self-driving cars In level 3 and 4 self-driving cars, drivers and automated driving systems(ADS) give and take control of their vehicles each other. In certain situations, such as traffic jam or long, straight, wide-open highway, if a driver wants to, the driver can choose self-driving mode and let the ADS drive the vehicle. However, when it encounters challenging road, traffic, or weather conditions, ADS signals the driver to take the wheel. The latter, taking control back from ADS, is called 'takeover'. Engineers and researchers agree that takeover is one of the most critical issues in developing and commercializing safer level 3 self-driving cars. For a smooth and safe transition, two conditions must be met. First, the signal for transition must be sent from ADS to the driver. This signal is called TOR(takeover request).[2] Second, TOR must not be too late. The time left to get into so dangerous situations after TOR, TORlt(takeover request lead-time), must be precisely calculated[2, 3]. If TORlt is too short, it is possible that the driver would not be in a position to manually drive the vehicle even when it is already in a situation where ADS cannot handle. This unstable and dangerous situation might be called 'the control gap'. To commercialize level 3 self-driving cars, the control gap must be closed. But closing the control gap is not just about commercialisibility. Leaving it open may seriously affect not only the driver's and the vehicle's safety but also others' and surroundings'. In this sense, closing the control gap raises normative concerns: to close the control gap, or to make takeover safe, what should ADS and the drivers do? This is the moral question that takeover raises.

To address this question, I have suggested the norms of coordination[4]. The norms of coordination consist of two norms: one for ADS and another for drivers. The norm for ADS is the norm of personalization, which states that ADS must be as personalized as possible for each driver. Why should ADS be personalized? As stated above, the crucial issue in safe takeover is calculating TORlt. TORlt is a sum over three time periods, driver's perception-response time for TOR(T1), mental processing time(T2), and settling time(T3). In all of these times, there are individual differences. For instance, disability and age may affect T1, T2, and T3. Further, driving skills and experiences would affect T2 and T3. But the most crucial factor is drivers' states. For a more detailed description of takeover and transition time, see Figure 1.[3]

[2] TOR usually takes a form of multimodal signals, for example, visual + sound signals.
[3] 'Driver readiness' means alertness or attention level.

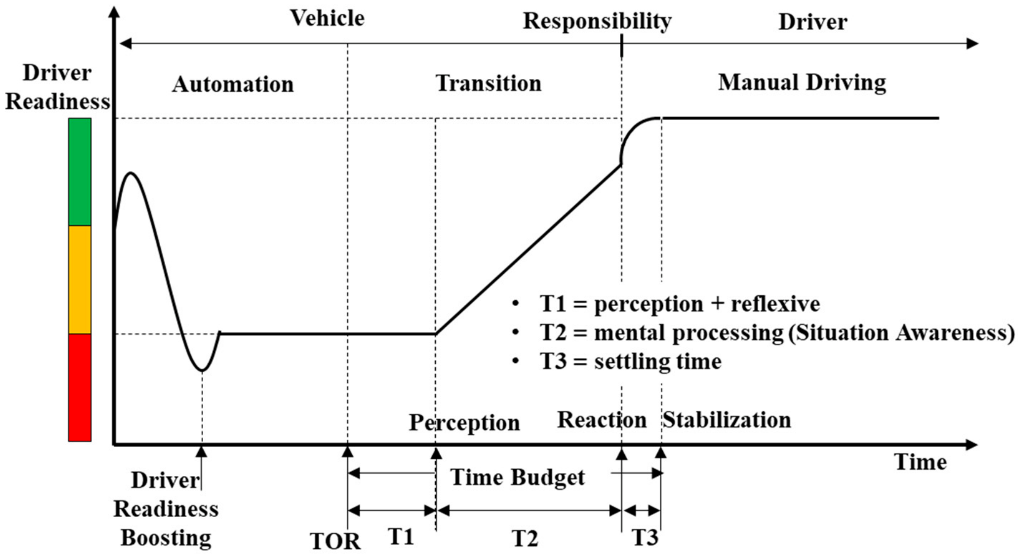


Figure 1. Concept of take-over transition time[5]

To calculate TORlt precisely, ADS must take account of how a driver is. Both physical and mental states are important. If a driver's head is turned to the back seats, it will definitely slow the driver's perception and reaction and affect T1. If the driver's attention level is low, it will affect T1~T3[6, 7, 8]. For precise TORlt calculation ADS must know detailed personal information about exactly in what state the driver is[9]. The norm for drivers is the norm of compliance, which states that drivers must comply to requests of ADS. Why should drivers comply to requests from ADS? To precisely calculate TORlt, ADS needs precise information about drivers. Providing false or unreliable information or even refusing to provide any information would necessarily disturb calculation of TORlt. From ADS's perspective, it is a sort of 'deception'. If a disabled driver does not comply to ADS's request of providing information about disability, such incompliance will make ADS miscalculate TORlt. The calculated TORlt may be actually too short for the driver, and this may open the control gap. Thus, for the sake of safe takeover, drivers must provide all the information ADS without any intended or unintended deceptions. That is, they must comply to requests of ADS. Both norms are not just about safety technology. They are ethical in nature, because any violation of them can cause serious harm to others' lives and properties. Also, both norms vividly show that the ethics of level 3 self-driving cuts both ways. ADS or drivers alone cannot guarantee safe takeover. To achieve safe takeover, both ADS and drivers should cooperate or coordinate with each other This is why they are called the norms of coordination. What should ADS and the drivers do for safe takeover? Coordination of personalization and compliance may be an answer to this question.

2. The norms of coordination vs data ethics: tension in the ethics of self-driving

Suggesting what should be done is one thing, and articulating how to do it is another. Drivers' compliance may be a matter of personal choice. ADS's personalization, however, is a matter of technology and design. As explained in Section 1, personalization of ADS is essentially about how to gather and process detailed information about drivers. And to gather and process information about drivers, the first thing ADS should do is data collection. Without data about a driver's state and driving activities, ADS cannot do anything. Collecting data about the driver and driving requires monitoring how the driver is and how the car is. One of the most crucial technical issues in developing and upgrading level 3 self-driving cars is driver monitoring technology. For example, if the driver is distracted by non-driving related-tasks(NDRT) or tired, ADS should know about this to calculate TORlt. Besides calculation of TORlt, data must be collected to prevent the driver from being 'out-of-the-roof', a situation where the driver cannot be in a position to perceive and react to TOR. If the driver is falling asleep, passing out, extremely unstable, or deluded, sending TOR or calculating TORlt would not work. The driver's state must be managed not to encounter such situations. Effective data collection for driver monitoring is the key to personalization of ADS and risk management as well. To successfully personalize ADS, data about drivers and driving activities must be effectively collected and processed.

What kind of data should be collected? Both drivers and their driving activities should be monitored. Most of all, biometric data seem to be necessary for driver monitoring. For instance, eye and head movement and direction of face would be crucial. They tell ADS where the driver's attention is headed and whether the driver falls asleep. Indeed, camera technology for eye tracking and facial recognition is very popular in the field. Body posture and gesture are also important, because they may affect the time needed for the driver to reach the wheel and step on pedals. This time partially determines the driver's perception-response time for TOR and settling time. In fact, in some models of level 3 self-driving cars, if drivers recline their seat over a certain degree in automated driving mode, alarm goes off. To monitor affective or emotional states of the driver, facial expressions should be recognized. Since emotion and affect are closely related to vital signs of body, heart rate, blood pressure, body temperature, and even EEG can be measured in real time. These data can be collected by built-in sensors or wearing devices. EEG is also important data for attention management for the driver. Further, driving data are required to monitor driving activities. Speed, distance between surrounding cars, acceleration and deceleration should be measured and stored in real time. Of course, ADS needs these data to calculate TORlt. These data are about driving activities and states of the car. But they also show states of the driver. ADS may infer how the driver is from how the car is. Sudden acceleration and deceleration can be a reliable sign that the driver is unstable. In short, monitoring drivers and their driving activities requires biometric and driving data.

Then, the norms of coordination can be reformulated in terms of data collection. The norm of personalization says that ADS must be as personalized as possible for drivers. We have seen that such personalization implies collecting data about drivers and their driving activities. Thus, to fulfill the norm of personalization, ADS must collect biometric and driving data. The same goes for the norm of compliance. The norms of compliance states that drivers must comply to requests of ADS. What would ADS request drivers? It would request certain personal information, such as disability and age. To monitor drivers and driving activities, ADS must collect biometric data of drivers and driving data of their cars. If so, to fulfill the norms of compliance, drivers must let ADS collect biometric data of themselves and driving data of their cars. The norm of personalization turns into the norm of data collection on drivers and their driving activities, and the norm of compliance becomes the norm of providing biometric and driving data to ADS. The norms of coordination are translated into 'the norms of data production'.

This is where ethics of self-driving and data ethics meet. For self-driving cars to run safely and ethically, a tremendous amount of biometric data must be collected through cameras, sensors, and wearable devices installed in the vehicle. ADS continuously collect data about how drivers are and how they drive. Collecting data about drivers and their activities, ADS learns a lot. It knows too much. ADS knows everything about where the drivers have been, what they did while driving, and how they drove their car. Considering the nature of the data collected by ADS, drivers may take such data collection as serious invasion of privacy or micro-surveillance on their body and behaviors. Once we choose automated driving mode, almost everything about our body and our car is collected, processed, and stored by ADS. Also, there are many security issues regarding data collection. Nothing is easier for self-driving car companies to use those data. The companies will know what ADS knows. Nobody can be sure what they will do with the data. Further, if there are data and information, there are hackers. The more data are collected and stored, the higher the risk of hacking increases. The data collected by ADS can be hacked and if placed under the wrong hands can lead to serious crimes. All the issues of data ethics are revived in the case of self-driving.

It must be noted that the problems mentioned above are unique to level 3 and 4. Technically, issues concerning biometric data occur only in the case of level 3 and level 4. This is because the need for biometric data collection comes from the normative concerns about takeover, and takeover only takes place in level 3 and 4 self-driving cars. No other level self-driving car involves such interaction. Level 1and 2 self-driving cars are just partially automated' by definition. There cannot be any automated driving mode in those levels. Since there is no automated driving mode, giving or taking control cannot occur either. In level 3 and 4, ADS has to monitor its driver's state because when the time comes, it has to return control of the car. But in level1 and 2, cars cannot take control in the first place. This is the reason why they usually are not taken as genuine self-driving cars. What about level 5? Level 5 self-driving cars are fully automated or driverless. Thus, the interaction between drivers and ADS is eliminated. And this is the reason why level 5 self-driving cars do not involve driver monitoring. There is no one to monitor.

Also, note that the problematic consequences are originated from normative concerns about takeover. For ethical level 3 and 4 self-driving, safe takeover must be achieved. To achieve safe takeover, ADS and drivers should follow norms of coordination. Close technical observations presented above, however, show that fulfilling norms of coordination implies to allowing data collection on drivers and their driving activities. Some may think allowing ADS to collect biometric and driving data is in itself unethical, because such data are so personal and sensitive that misuse of them could do serious harm to others' rights, properties, and lives. Can we really accept these consequences of norms of coordination? If we cannot, we have to admit that ADS and drivers may refuse to follow norms of coordination. Nonetheless, without those norms, there seems to be no way to handle the risk of the control gap. As emphasized in Section 1, reducing the risk of the control gap to make takeover safe is a normative matter. When it comes to level 3 or 4 cars, ADS must reflect individual differences and drivers must comply with requests of ADS. They are like duties for ADS and drivers to follow. Not fulfilling those duties seems unethical, since it is to risk others' lives and properties for one's privacy and security. In so far as ADS and drivers are morally obliged to prevent the control gap, it is inevitable to fulfill norms of coordination. It is normative and ethical concerns that takeover enforces norms of coordination, and norms of coordination do not go well with those of personal data protection.

The technology of takeover should be taken seriously. To know how ethical self-driving is entangled with data ethics, one must understand how takeover actually works. If there was any technically possible way for ADS to monitor drivers' states without collecting biometric data, and if monitoring driving activities did not require driving data, data ethics would not matter. However, monitoring without data appears to be technologically impossible. Without biometric data, ADS gathers no information about drivers' states. Collecting, processing, and storing driving data is the only way to have information about driving activities. As takeover necessarily involves data, data ethics always come in. Ethical self-driving will be a matter of data ethics, because takeover cannot be done without data collection.

If the argument so far is on a right track, there is tension in ethics of self-driving. On the one hand, the attempt to make self-driving ethical necessarily requires data collection and easily turns into micro-surveillance. On the other hand, it is highly likely that ethical concerns about data collection would hold back progress on the achievement of safe takeover and prevent self-driving from being ethical. For safe takeover and ethical self-driving, the norms of coordination must be fulfilled. However, the technology required in fulfilling those norms inevitably raises the risk of unethical use of personal data. For personal data protection, data collection should not be allowed. Or, it must be legally regulated at least. Banning or regulating data collection is likely to increase the risk of the control gap. To avoid the risk of the control gap, one should raise that of micro-surveillance and might face many security threats. To avoid the risk of micro-surveillance, one should raise that of the control gap by giving up making takeover safe.

3. Data ethics for self-driving: what is going on and what has to be done

Even though there is tension between ethical self-driving and data ethics, it does not have to be seen as an unsolvable dilemma. Self-driving is not merely an example for thought experiments or ethical debates. It is a pressing real issue. Thus, we have to assess the tension in ethics of self-driving not only ethically but also realistically. One may find a way to ease that tension by considering what is going on and what has to be done. What is actually going on in the self-driving industry tells us what has to be done for ethics of self-driving.

Given that various driver monitoring technologies are already used in self-driving vehicles, banning data collection from self-driving cars is unrealistic. Engineers and researchers in the self-driving industry are developing software and hardware for more comfortable and safe self-driving cars. Without data about drivers and their driving activities, nothing can be done in safe takeover. And without safe takeover, no one would use automated driving mode. In this sense, to ban data collection is just to ban self-driving cars. Further, data collection may provide extremely valuable resources for more efficient and safer self-driving vehicles. The more data are collected, the faster self-driving cars evolve. If so, banning data collection might be giving up the valuable opportunity to make better self-driving cars. As noted in the beginning of this article, we are living in the age of self-driving. People already drive self-driving cars and allow ADS to gather their data. All things considered, banning data collection seems anachronism.

Data collected from self-driving cars can be used for various purposes. Yet, this powerful resource comes with challenges. To meet these challenges, the most important thing would be data ethics tailor-made for self-driving. Takeover is in itself a matter of data ethics. Those who are involved in the development and production of self-driving cars, such as engineers, programmers, and data scientists, are likely to interact with data about drivers and their driving activities daily. Data ethics are of the utmost concern to them. They must be well-versed in basic concepts and principles of data ethics, which encompass the moral obligations of gathering, protecting, and using personal data. While they may not be directly responsible for designing or implementing code, managing a database, or training a machine-learning algorithm of ADS, understanding data ethics can help them to detect any unethical collection, storage, or use of drivers' data. Catching unethical cases is crucial not only for protecting drivers' data but also for saving companies from legal issues. To this end, people in the self-driving industry should be familiar with general issues in data ethics, consent, transparency, privacy, etc. By adapting and translating data ethics into technical contexts of self-driving, they will have data ethics for self-driving. And based on data ethics for self-driving, data privacy laws concerning data collection in self-driving vehicles can be legislated. Since developing self-driving cars always involves intensive market competition, merely establishing and training a new field of ethics will not work. To properly and effectively regulate the use of data, laws concerning data collection must be legislated. Well-established data ethics and data privacy laws, once they are well-established, will effectively resolve the tension between ethical self-driving and data ethics.

References

1. SAE. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, https://www.sae.org/standards/content/j3016_202104/ (accessed April 15 2021)
2. M. R. Endsley and E. O. Kiris, The out-of-the-loop performance problem and level of control in automation. *Hum. Factors*. **37**, 2 (1995)
3. C. D. Wickens, J. G. Hollands, S. Banbury and R. Parasuraman, *Engineering psychology and human performance,* Psychology Press, (2015)
4. Moon, K. Takeover as an Ethical Problem: Level 3 automated driving car and norms of coordination. *Philosophy·Thought·Culture*. **38** (2022)
5. Kim, J. Kim, H.-S. Kim, and W. Yoon, D, Take-over performance analysis depending on the drivers' non-driving secondary tasks in automated vehicles. *Proceedings of the 2018 International Conference on Information and Communication Technology Convergence* (2018)
6. S. Bao and L. N. Boyle, Age-related differences in visual scanning at median-divided highway intersections in rural areas. *Accid Anal Prev*. **41**, 1 (2009)
7. H. Jeong, Z. Kang & Y. Liu, Driver glance behaviors and scanning patterns: Applying static and dynamic glance measures to the analysis of curve driving with secondary tasks. *Hum. Factors Ergon. Manuf. Serv.* 29 (2019)
8. G. H. Robinson, D. J. Erickson, G. L. Thurston and R. L. Clark, Visual search by automobile drivers, *Hum. Factors.* **14**, 4 (1972)
9. K. Zeeb, A. Buchner and M. Schrauf, What Determines the Take-over Time? An Integrated Model Approach of Driver Take-over after Automated Driving. *Accid Anal Prev*. **78** (2015)

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.021

USER GUIDED AUGMENTED REALITY MANUFACTURING ASSEMBLY INSTRUCTION DEVELOPMENT: CHANGING WORK MODES DURING COVID -19 AND ITS IMPLICATIONS

IVETA EIMONTAITE
School of Aerospace, Transport and Manufacturing, Cranfield University, UK
E-mail: iveta.eimontaite@cranfield.ac.uk

THODORIS TOGIAS
Laboratory for Manufacturing Systems and Automation, Department of Mechanical Engineering and Aeronautics, University of Patras, Greece
E-mail: togias@lms.mech.upatras.gr

SARAH FLETCHER
School of Aerospace, Transport and Manufacturing, Cranfield University, UK
E-mail: s.r.fletcher@cranfield.ac.uk

NIKOS DIMITROPOULOS
Laboratory for Manufacturing Systems and Automation, Department of Mechanical Engineering and Aeronautics, University of Patras, Greece
E-mail: dimitropoulos@lms.mech.upatras.gr

Robots and automated technology are still being perceived as threatening by operators in manufacturing. The current study investigates how participants' feedback and evaluation of the user interfaces guide the final design. The initial development of the Augmented Reality (AR) guided assembly was completed over several stages; however, the current study focuses on testing the product after the initial comments on the AR demonstrator from the users were integrated. The work discussed here was completed over three steps: (i) An online survey where participants (N = 21) viewed pre-recorded videos of an experimenter completing the task and answered questions related to clarity and usability of the AR glasses; (ii) A trial experience where participants (N = 5) were invited to the lab to complete the task using the proposed solution, and (iii) experimental study where participants (N = 10) interacted with a robot via the developed AR guided instructions in two conditions where robot communicated its next steps (the transparency condition) and where it did not communicate its next steps (the neutral condition). The comparison between step (i) and (ii) provides information not only between two development points of the AR glasses guided instructions, but also estimates how at hand experience (trial experience and interview) influences the perception of the technology compared to distance experience (online survey). Finally, the experimental study indicates that robot behaviour transparency can affect experienced lower physical demand when the user knows the robot's next steps. The current paper will address issues of the development and introduction of technology while incorporating the users.

Keywords: human-robot collaboration, mental workload, user-centred design, trust, situation awareness, augmented reality guided assembly

1. Introduction

The introduction of robotic technology in manufacturing can be a challenging task from a technological and a human factors point of view. Technology rejection by the user is enhanced by lack of trust, which, among other reasons, can be caused by a lack of transparency in robot behaviour [1]–[3]. Yet, the understanding of the way people work with this new technology is still unclear with limited research done on the topic. Therefore, the communication interfaces may increase acceptance and trust on new technologies [4]–[7].

Lack of trust in new technology, rejection of new processes, and decreased motivation are consequences of psychological strain operators' experience if human factors are not properly accounted for when introducing a new technology. Being able to experience the technology, and a transparent introduction to the robot capabilities creates realistic expectations of this technology [2], [8]–[10]. This also can affect the success of introduction of technology and acceptance of new processes [11]. Furthermore, transparency in the technological capabilities and the establishment of clear rules and responsibilities would help to ease the uncertainty within new processes [12], [13].

On the other hand, robot's communication of its behaviour might cause ambiguity and decrease the sense of safety for the operator. Information communication in a clear and concise way leads to psychologically and physically healthier and safer environment [14]–[18]. It should be intuitive and well defined how an operator communicates with the robot, and how the robot would communicate about its next moves and directions. Moreover, robots' behaviour has a great influence on building trust [10], and when the operators do not trust the robot they tend to monitor its performance more than needed. In such cases, mental workload and cognitive demand increase due to monitoring robot performance. Such increase results in less cognitive capacity left for monitoring the environment and complying with safety procedures [6], [7]. Finally, lack of trust and increased monitoring of the robot might relate to decreased situation awareness. Situation awareness is necessary in order for the individual to operate in the environment safely, while a decrease in situation awareness can lead to an increase in work accidents and injuries [19], [20].

The main barriers for the technology acceptance and engagement by the user is the delayed introduction and lack of involvement in the development stage. The current study, conducted as part of the SHERLOCK project, investigates how participants' feedback and evaluation of the user interfaces guides the final design and affect users trust in industrial human-robot collaboration, situation awareness and mental workload. To involve potential users of Augmented Reality (AR) guided instructions for manufacturing and human robot collaboration tasks, the study was done in two stages. Firstly, the developed AR instructions were evaluated by the participants. The changes in working patterns due to Covid-19 over the last two years suggest that technology introduction and initial training methods will change, therefore, this step was conducted comparing online introduction to this technology with hands-on experience in the lab. Secondly, a human robot collaboration experimental study was conducted investigating the impact of robot transparency (in this case communication of its next steps) on user trust, situation awareness and mental workload.

2. Study 1 Methods

2.1. *Design*

The main aim of online study and trial experience was to compare the participant evaluation and feedback on the instructions on how to complete the manufacturing assembly tasks. The online study provided first person view of the AR glasses while completing the task, and in the other one participants completed the tasks themselves in the lab.

2.2. *Participants*

Twenty-one participants took part in online study. Six participants were manufacturing employees and 15 were university students and staff. Due to anonymous nature of the study, it was not possible to verify which use cases manufacturing employees came from. The average age of participants were 30.65 years old (SD = 5.97, min = 24, max = 46). 40% of participants had no experience with AR technology, while 60 % had "Some but never on a regular basis".

The hands-on trial experience the group consisted of five participants. Out of five participants, three were males, one female, and one identified themselves as other. Four of the participants were working at the University of Patras, and one participant was working both at the university and the manufacturing industry. Participants were asked about their experience with robots and AR devices; three participants indicated daily experience with robots, and two

- once a week; furthermore, four participants had experience with AR devices before, while one did not have any experience before.

Ethics approval was received form Cranfield University and University of Patras Ethics Committees.

2.3. *Materials*

Qualitative and quantitative questions were used to evaluate user experience of completing the assembly task with AR instructions. Qualitative questions asked for participants evaluation and feedback on instructions presented via AR glasses, voice commands, user interface and 3D animations (i.e. "what did you like about AR instruction presentation", "what would you change in the AR instruction presentation"). This data was used for further improvement of the AR guided instructions and human-robot interaction on the large panel assembly task and is not going to be presented in this study. The quantitative aspect was focusing to assess participants' evaluation of instructions Statements such as "*I found the instructions presented via VR headset to be effective/simple/precise/understandable/concise/predictable*" were asked to be rated from -50 (disagree with a statement (i.e. intrusions were not effective)), 0 neutral opinion (neither effective, nor ineffective) to 50 (agreeing with the statement (very effective)). With the respect of voice commands, participants were asked whether voice commands were self-explanatory and easy to memorize. Finally, the user interface aspect was concerned if it looks and feels friendly and whether it provides quick access to features and commands. All these aspects were evaluated on the scale from -50 (disagreeing with the statement), 0 (neither agreeing nor disagreeing), and 50 (agreeing with the statement).

2.4. *Apparatus*

AR glasses Microsoft Hololens 2 was used to guide operators on large industrial panel assembly. The instructions provided information about the necessary tools, sequence of steps and overlayed position on the panel where participants needed to assemble the components. Gesture (hand-detection) and voicebased commands were enabled to allow participants to go through the instruction steps

2.5. *Video clip*

Online context participants were provided with a video depicting the assembly task guided by AR instructions in collaboration high payload collaborative robot (AURA collaborative robot, Comau, see section 4.4). The video lasted for 3 min 38 seconds and can be viewed here: https://www.youtube.com/watch?v=mnI9OUjplR4.

2.6. *Procedure*

Once agreeing to take part, participants were provided with the information about the study and informed consent on Qualtrics the online platform. Online study participants were presented with the video of first-person perspective depicting the assembly of large component with the robot while using AR instructions for assembly guidance. On-hand trial participants after signing the informed consent were asked to complete the task in the laboratory. Participants were assisted with putting on the AR glasses, given a short demonstration of how to navigate the AR environment and asked to complete the task. Once they completed the assembly task, they were asked to answer the same questions as the online study participants on the Qualtrics platform. As a final step, participants from both studies were debriefed.

2.7. *Study 1 Results*

Participants from both the online survey and interview session were asked to provide the quantitative scores on their experience of instructions, voice commands and user interface. In this section the comparison between these two sessions will be conducted providing information not only between two development points of the AR glasses guided instructions, but also estimate how hands-on experience (trial experience and interview) influenced the perception of the technology compared to distance experience (online survey).

The adjectives describing participants attitudes to the instructions indicate that they all perceived them positively (Fig 1). Participants indicated that instruction effectiveness, understandability, and simplicity where the most positive aspects after real life experience, while online survey participants indicated that the most positive aspects of the instructions were the fact that they were concise, effective and simple. Further analysis with Mann-Whitney U test indicated that real experience participants scores were significantly higher on statements relating to “simple”, “effective”, “understandable” and “predictable” aspects of the instructions ($U = 9.50$, $p = .011$, $U = 8.00$, $p = .010$, $U = 10.50$, $p = .015$ and $U = 7.00$, $p = .019$, respectively). However, these differences were not significant after adjusting with Bonferroni correction for multiple comparison (2-tailed alpha = .008).

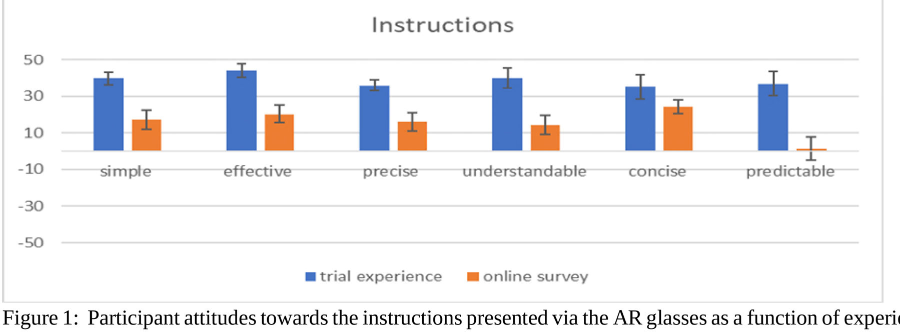


Figure 1: Participant attitudes towards the instructions presented via the AR glasses as a function of experience modality (real trial vs. online)

The adjectives describing participants attitudes to the voice commands and user interface indicate that they all perceived them positively (Fig 2). Although there was no significant difference between voice commands and participant group ($p \geq .387$), participants’ responses differed regarding attitudes towards interface: “looks and feels friendly” and “provides quick access to features and commands” were both evaluated more positively in the trial experience condition ($U = 11.50$, $p = .035$ and $U = 10.00$, $p = .026$). Yet, only “provides quick access to features and commands” was at a trend significance after Bonferroni correction (2-tailed alpha = .025).

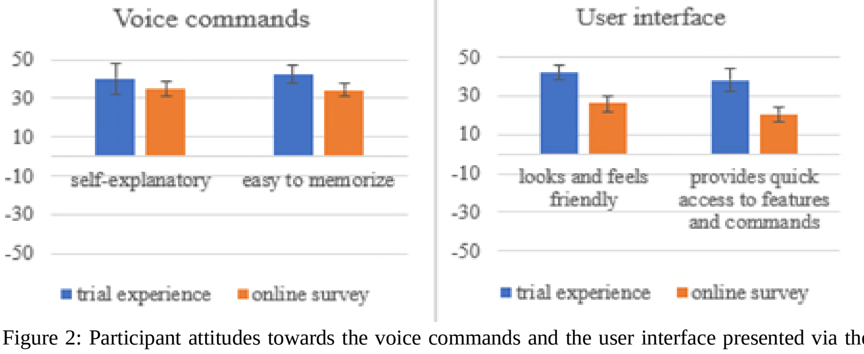


Figure 2: Participant attitudes towards the voice commands and the user interface presented via the AR glasses as a function of experience modality (real trial vs. online)

3. Study 2 Methods

3.1. *Design*

The experimental study investigated how the developed and adjusted AR guided assembly instructions can be used in the industrial human-robot collaboration. The main aim of the study was to capture cognitive and human factors involved while working with a high payload Aura robot as a function of the experimental condition (transparency where robot next moves are projected visually vs. no transparency).

3.2. *Participants*

Ten participants (1 female) with the mean age of 25.4 years (SD = 0.81) took part in the study. All participants were working at the University of Patras, and five of them had daily experience with robots (one-"once a week", three "sometimes, but never on a regular basis"). Finally, when asked about experience with the AR technology, three participants indicated that they "never" interacted with it, five "sometimes, but not on a regular basis", and two participants indicated that they experience this technology "on a daily basis". The study received ethics approval from Cranfield University (UK) and the University of Patras (Greece).

3.3. *Materials*

The study assessed participants mental workload, situation awareness and trust in industrial human-robot collaboration via the established and validated questionnaires. Mental workload was measured using the NASA Task Load Index (NASA-TLX; [21]. It measures participants' experience of a task across six different dimensions such as physical demand, frustration level and performance. Participants' trust was measured using the Trust in Industrial Human-Robot Collaboration scale [22] and situational awareness was measured with Situational Awareness scale [23].

3.4. *Robot*

For the industrial human robot collaboration task, a high payload collaborative robot AURA (Comau, S.p.A, Italy) was used to assist participants with rotation of the large industrial panel to assist on the assembly. This robot has six safety levels to support safe human collaboration while performing manual tasks without the need for barriers or fences.

3.5. *Procedure*

The study was conducted in the LMS laboratory with laboratory members. The experiment was completed in a single session lasting around 30 min. Upon coming to the testing area, participants were provided with informed consent, informed about their participant rights, information about the task and experiment, how their data will be stored and used, and provided an opportunity to ask any questions relating to the study. Participants were additionally briefed about safety measures during the experiment relating to Covid -19. After they have signed the informed consent, the experiment began.

Participants were assisted in putting on AR glasses, explained the task, and performed half of the assembly followed by filling in the questionnaires. After the first set of questionnaires were filled in, participants finished the assembly task. All participants completed the task with the robot in two different conditions in the counterbalanced order. After the last condition was completed and the questionnaires filled in, participants completed demographic information (age, gender, experience with robots, experience with AR devices). As a final step, participants were debriefed.

4. Results

4.1. *Mental workload*

Mental workload was measured using the NASA-TLX. Non-parametric tests were used for the inferential statistics to account for the low participant number. As a first step, overall scores for each dimension was compared. This analysis yielded significant results between all dimensions with frustration dimension (Table 1).

Table 1: Inferential statistics for mental workload dimension comparison for both conditions combined

	Physical demand	Temporal demand	Effort	Frustration	Performance
Mental demand	Z = -0.89, p = .374	Z = -0.26, p = .799	Z = -.56, p = .676	Z = -2.55, p = .011	Z = -0.28, p = .770
Physical demand		Z = -0.51, p = .610	Z = -.56, p = .575	Z = -2.55, p = .011	Z = -0.97, p = .333
Temporal demand			Z = -1.27, p = .203	Z = -2.81, p = .005	Z = -0.255, p = .799
Effort				Z = -2.70, p = .007	Z = -0.89, p = .373
Frustration					Z = -2.60, p = .009

Further exploration between the conditions showed that there was a significant difference between physical dimension ($Z = 2.37$, $p = .018$) where in the transparency condition participants felt less physical demand than in no-transparency condition (Fig 3).

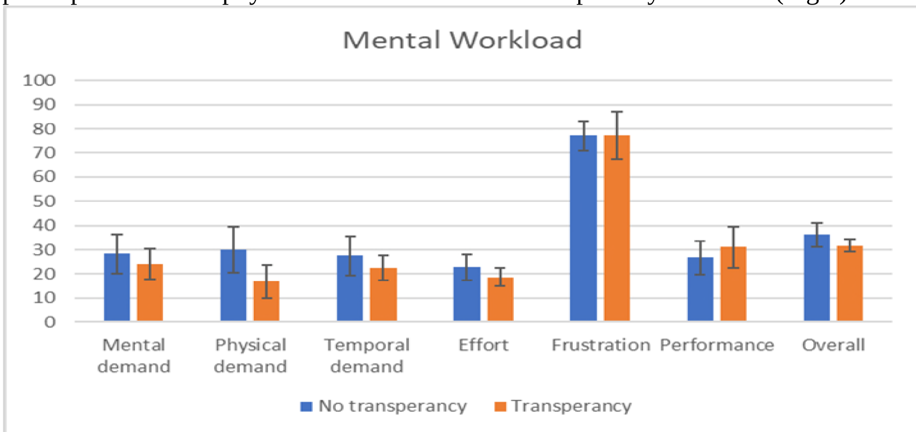


Figure3: Mean (+/- SEM) across mental workload dimensions as a function of experimental condition

4.2. *Trust and Situation Awareness*

Responses for the Trust in Industrial Human-Robot Collaboration scale and Situational Awareness scale indicate that there was little difference between responses after transparency and no-transparency conditions ($p \geq .05$). The future analysis will investigate between subject differences on participants who had at first transparency vs no-transparency trials.

5. Discussion

The current paper reports the work conducted to provide user evaluations and feedback on the development of the development of AR guided Human-Robot Collaboration instructions and the impact of robot transparency on participants situation awareness, trust and mental workload. This is particularly relevant in fields were user acceptance and trust in technology is of high importance.

The results indicate that participants had positive attitudes towards the AR technology both after viewing the videos and interacting in real life. However, it is important to note, that the comparison of the evaluation scores shows that hands-on experience participants had more positive attitudes towards the AR assisted assembly guidance. This is important because of two aspects. Firstly, due to Covid-19 pandemic there is an increased need for the remote work, and therefore there is a possibility that in future remote training might become more common. The results suggest that capturing feedback via remote means provide consistent results with real life training. Secondly, having higher scores on the real-life experience provides further emphasis on the importance of operator engagement in the early stages of technology development and introduction. Allowing individuals to experience the technology on the existing tasks yield more positive attitudes towards new technology.

Furthermore, the results of the experiment indicate that a high payload collaborative robot in transparency mode can alleviate self-reports of physical demand and increase self-perceived performance scores. Interestingly, the frustration dimension was significantly higher than other dimensions (e.g. mental, physical and temporal demands). One possible explanation being that task clarity or waiting times for the robot might have influenced such results. However, this result was unexpected and calls for further analysis: It is planned that further analysis will investigate the eye gaze patterns and correlate with the self-reported data to explore this unexpected result.

Automation role and need in the manufacturing industry is increasing significantly [24]. Yet, the lack of work done to assess an operator psychological safety can be detrimental in the integration of robotic applications [25]. Using a bottom-up approach, engaging users early on and increasing operators' ownership of new processes can positively affect employees' wellbeing and cooperation in further technological development [26], [27]. The findings and observations reported in this work serve not only to direct technology development, but also contributes to the development of methodology on how to assess operator psychological safety and human factors with the final SHERLOCK solution demonstrators. The results will feed into the final SHERLOCK studies while integrating the developed technology on the use case companies’ shopfloor.

Acknowledgments

This work has received funding from the European Union Horizon 2020 research and innovation programme as part of the project SHERLOCK under grant agreement No 820689

References

[1] B. M. Muir, ‘Trust between humans and machines, and the design of decision aids’, *International Journal of Man-Machine Studies*, vol. 27, no. 5–6, pp. 527–539, 1987.

[2] T. L. Sanders, T. Wixon, K. E. Schafer, J. Y. C. Chen, and P. A. Hancock, ‘The influence of modality and transparency on trust in human-robot interaction’, in *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, San Antonio, TX, USA, Mar. 2014, pp. 156–159.

[3] R. H. Wortham and A. Theodorou, ‘Robot transparency, trust and utility’, *Connection Science*, vol. 29, no. 3, pp. 242–248, Jul. 2017, doi: 10.1080/09540091.2017.1313816.

[4] A. R. Selkowitz, S. G. Lakhmani, C. N. Larios, and J. Y. C. Chen, ‘Agent Transparency and the Autonomous Squad Member’, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, no. 1, pp. 1319–1323, Sep. 2016.

[5] S. Ososky, T. Sanders, F. Jentsch, P. Hancock, and J. Y. C. Chen, ‘Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems’, in

Unmanned Systems Technology Xvi, vol. 9084, R. E. Karlsen, D. W. Gage, C. M. Shoemaker, and G. R. Gerhart, Eds. 2014, p. 90840E.
[6] J. Chen, B. W. Ren, X. Song, and X. Luo, ‘Revealing the “Invisible Gorilla” in Construction: Assessing Mental Workload through Time-Frequency Analysis’, 2015.
[7] D. J. Saxby, G. Matthews, J. S. Warm, E. M. Hitchcock, and C. Neubauer, ‘Active and Passive Fatigue in Simulated Driving: Discriminating Styles of Workload Regulation and Their Safety Impacts’, *J Exp Psychol Appl*, vol. 19, no. 4, pp. 287–300, Dec. 2013,
[8] M. W. Boyce, J. Y. C. Chen, A. R. Selkowitz, and S. G. Lakhmani, ‘Effects of Agent Transparency on Operator Trust’, in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, New York, NY, USA, 2015, pp. 179–180.
[9] J. Y. C. Chen, M. J. Barnes, A. R. Selkowitz, K. Stowers, S. G. Lakhmani, and N. Kasdaglis, ‘Human-Autonomy Teaming and Agent Transparency’, in *Companion Publication of the 21st International Conference on Intelligent User Interfaces*, New York, NY, USA, 2016, pp. 28–31.
[10] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, ‘A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction’, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, Oct. 2011.
[11] R. D. Iverson, ‘Employee acceptance of organizational change:the role of organizational commitment’, *The International Journal of Human Resource Management*, vol. 7, no. 1, pp. 122–149, Feb. 1996.
[12] K. Schaefer, J. Chen, J. Szalma, and P. Hancock, ‘A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems.’, *Human factors*, 2016.
[13] C. D. Wickens, J. G. Hollands, S. Banbury, and R. Parasuraman, *Engineering Psychology and Human Performance Christopher D. Wickens, Justin G. Hollands, Simon Banbury, Raja Parasuraman*. Psychology Press, 2015.
[14] G. Bahar, M. Masliah, R. Wolff, and P. Park, ‘Desktop reference for crash reduction factors’, 2007.
[15] K. R. Laughery, ‘Safety communications: Warnings’, *Applied Ergonomics*, vol. 37, no. 4, pp. 467–478, Jul. 2006.
[16] M. E. Mills and K. Sullivan, ‘The importance of information giving for patients newly diagnosed with cancer: a review of the literature’, *Journal of clinical nursing*, vol. 8, no. 6, pp. 631–642, 1999.
[17] C.-H. Tang, W.-T. Wu, and C.-Y. Lin, ‘Using virtual reality to determine how emergency signs facilitate way-finding’, *Applied Ergonomics*, vol. 40, no. 4, pp. 722–730, Jul. 2009.
[18] E. Vilar, F. Rebelo, and P. Noriega, ‘Indoor Human Wayfinding Performance Using Vertical and Horizontal Signage in Virtual Reality: Indoor Human Wayfinding and Virtual Reality’, *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 24, no. 6, pp. 601–615, Nov. 2014.
[19] M. R. Endsley and M. M. Robertson, ‘Training for situation awareness in individuals and teams’, in *Situation awareness analysis and measurement*, NJ: Erlbaum Mahwah, 2000, pp. 349–366.
[20] A. Sneddon, K. Mearns, and R. Flin, ‘Situation awareness and safety in offshore drill crews’, *Cogn Tech Work*, vol. 8, no. 4, pp. 255–267, Oct. 2006.
[21] S. G. Hart and L. E. Staveland, ‘Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research’, in *Advances in Psychology*, vol. 52, Elsevier, 1988, pp. 139–183.
[22] G. Charalambous, S. Fletcher, and P. Webb, ‘The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration’, *International Journal of Social Robotics*, 2016.
[23] R. M. Taylor and S. J. Selcon, ‘Cognitive Quality and Situational Awareness with Advanced Aircraft Attitude Displays’, *Proceedings of the Human Factors Society Annual Meeting*, vol. 34, no. 1, pp. 26–30, Oct. 1990.
[24] J. Manyika ***et al.***, ‘A Future that Works: Automation, Employment, and Productivity. McKinsey Global Institute.’, McKinsey Global Institute, 2017.
[25] M. D. Coovert and L. F. Thompson, *The Psychology of Workplace Technology*. Routledge, 2013.
[26] S. T. Hunter, K. E. Bedell, and M. D. Mumford, ‘Climate for Creativity: A Quantitative Review’, *Creativity Research Journal*, vol. 19, no. 1, pp. 69–90, May 2007,.
[27] M. Lukes and U. Stephan, ‘Measuring employee innovation: a review of existing scales and the development of the innovative behavior and innovation support inventories across cultures’, *International Journal of Entrepreneurial Behavior & Research*, vol. 23, no. 1, pp. 136–158, 2017.

SECTION–6

ROBOTICS AND ARTIFICIAL INTELLIGENCE

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.011

THE DIRECTION OF ARTIFICIAL INTELLIGENCE EDUCATION IN THE SCHOOL EDUCATION FIELD INTRODUCTION

TAECHANG KIM
Seoul Myunnam Elementary school, Seoul, Korea
dukeofslime@naver.com

This article is talking about the possibility that artificial intelligence development may cause various problems in the society. It is claiming about the necessity of focusing on AI-ethics in school curriculums, rather than its techniques. AI-ethics is completely different from practical ethics, but it is related and able to affect every people and technologies, hence AI ethics must be taught in schools. Although there are many tries to put AI-ethics classes in the curriculum, there is a problem; the uncertain definition of AI-ethics education; What is AI-ethics, what constituents are needed to be in the school curriculums, and what is the reason to teach it. This article divides the ethics – the ethics of AI, the ethics to treat AI, and the ethics about AI. And it clarify what AI ethics should be taught in schools and why. Also it insists the imperativeness of the staged AI-ethics education starting from elementary schools.

1. Introduction

Every year, the new technologies appears. AI is being used in almost everywhere. It has started to be used not only in the areas where the people easily spot, like robots or meta-verse, but also in everyday products and programs around; such as smartphones, computers, games, and refrigerators. If AI is in everyday life, school curriculum should contains the classes which teaches about AI. Then, what AI contents should be in the school curriculums? Many countries are introducing AI-related classes into school curriculums. Learning the use and technology of AI is important, but understanding how to use AI in a proper way is more important. If people abuse AI, the negative effect can reach to thousands of people. With current technology, it is not able to expect AI to have human’s ethics. The more AI being used, the more people who use AI will be asked of the ethics to use AI. Therefore, AI-ethics education needs to be started from an early age, and educate people about the guide line of using AI morally.

2. Artificial intelligence ethics education in Korea

Korea stipulates coding & software education as the compulsory education since 2016. Korean government have provided AI-teaching training to the teachers, and have supplied tools and materials to schools for AI classes. Although exquisite AI computers or devices haven’t been distributed or utilized at school sites yet, AI-related education is being conducted in diverse ways. However, in Korea's current national level curriculum, the 2015 revised curriculum, there are no specific AI-related content elements, AI-educational methods, or AI-learning achievement standards. Also in the curriculum, there is no system of AI-ethics education contents (Hyeongbin Park, 2020, 307). The elementary school moral subject textbook, which is based on this curriculum, is not containing AI-related contents or AI-ethics contents(Hamin Kim, 2021, 2). The lack of the systematic AI-education caused improvisation at the school sites, and it linked to the result; the AI ethics classes focused on vocational contents.

In『The Future of Jobs Report 2018』,released by the World Economic Forum (WEF) in 2018, says lots of jobs are disappearing and appearing at the moment, and the change is expected to be greater in the future (World Economic Forum, 2018). It also says that the careers and occupations will be changed – the jobs that were popular in the industrial and digital society may no longer be big, and completely new professions may appear, or the occupations that have been neglected may become huge. With these hypotheses, it is difficult to predict the future professions or students' career paths, which are expected to undergo rapid changes. Thus, the schools needed a new approach to the career education, different from the existing one. First, the coding and software education, which is considered as the basics of AI use, were introduced through the practical art subjects. However, the experts' opinions on the development level of future technology and occupations are not consensus. The experts overall believe that many jobs will disappear and lots of new jobs will emerge compared to the past. Although, they did not coincide in opinion about the jobs that appear as new or are treated better than the past. There are two representative arguments; whether AI will ultimately outperform humans in the workplace or not(Alex Ross, 2016, 53). The schools cannot immediately follow the changes described so far, so there is limitation in the curriculum. Current AI education in the school curriculums is for developing AI, based on computer science or technology, such as coding lessons or software education, and for vocational training (Taechang Kim, Soonyong Byun, 2021, 82).

3. The direction of artificial intelligence education needed in school education

What should we teach in the classroom when we teach AI? If AI is taught in the schools, it is because of the thought that it can help students' lives and lead to righteous lives. In the 2022 revised Korean curriculum, it will be revised to cultivate AI literacy by linking and operating the basic principles of AI and the use of AI (Ministry of Education, 2020, 30). However, in the current Korean curriculum, the ultimate purpose of the AI education is not clear – Is it 'for having the students become an AI inventor' or 'for raising the students to use AI effectively' or 'for letting the students utilize AI uprightly'? Furthermore, even though Korean schools have been established lots of classes in various way to strengthen AI education, the classes were struggled to set a target; learning specifically about AI, or learning how to handle AI. Thus at the current school sites, the students are focused on learning skills – understanding the use cases or the concepts of artificial intelligence, and the application of machine learning. The following table is an excerpt from Korea's data surveyed by the KFSC on AI curriculum according to the school levels(Korean Foundation for Science and Creativity 2020(a), 9) (Korea Foundation for Science and Creativity, 2020(Korea Foundation for Science and Creativity, 2020) b), 8).

Table 1. AI curriculum and educational materials by school level in Korea

	Elementary School	Middle School	High school
Korea	[Grades 3-4] ·Use cases of AI (Voice recognition, etc.) [Grades 5-6] ·Understanding the concept of AI ·Utilization of AI (Machine learning, etc.)	[Grade 1] ·Understanding AI knowledge expression (If then, Tree) ·Understanding the machine learning [Grade 2] ·Understanding AI inference methods ·Understanding machine learning approaches (Supervised, Unsupervised, Reinforcement, etc.) [Grade 3] ·Inference of knowledge in various fields ·AI learning in various fields (Domain)	[Grade 1] · Machine learning concepts and algorithms ·Data analysis method of machine learning (Probability, Statistics) [Grade 2] ·Concept and program of artificial neural network ·Application of artificial neural network to real life problems

As shown in the table above, AI education in the curriculum exists only to develop computational thinking ability, and to raise AI developers through coding education and software education. Namely, the curriculum only concentrates on having the students enable computational thinking through teaching computer languages and programming, just as the industrial society raises students to be able to use computers. The current AI education in Korea can be said as a vocational education that teaches the skills to build AI, and by extension, it is drawn to the expectation that all students will become AI experts (Kim Tae-Chang, Soonyong Byun, 2021, 83).

Here, I insist of the necessity of AI-ethics training, just as the importance of the AI-skill training. In the future society, AI will be essential. However, that doesn't mean that everyone has to be an AI expert who can create AI and code programs. Even when the digital society first started, it was difficult to run a computer without an operating system. As the technology advanced, the GUI operating system is generalized. The web browsers and the application programs also became popularized, the most people can handle the computers without understanding the computer languages or the operating system of the CUI system. Likewise, people will be able to use AI easier and more convenient someday. Even in the digital age, it has become more important to the ordinary citizens, except experts, to use the program correctly and not harm people with digital technology rather than how to make the program. Because the digital-ethics education was delayed in the digital society, numerous side effects appeared in the digital world. A similar thing could happen in a future society where AI becomes commonplace. Therefore, although the official AI-education has not occurred yet, I agree with the argument that ethical education is necessary before the side effects able to be caused by AI appearance (Soon-yong Byun, 2020, 428), and I also assert that we need to find out what and how AI-ethics should be taught in the curriculum.

4. The direction of AI ethics education

As mentioned in the previous chapter, if AI education is focus on the technology parts, then the purpose will be having the students be able to create and develop AI or the software, without considering the students' aptitude. Perhaps in the future, lots of jobs will be related to AI, but that doesn't mean everyone will be involved with building AI. Through the industrial and digital revolution era, scientific technology developed rapidly and became the center of the society, but not everyone was required to become a scientist or an engineer. In the industrial age, not everyone had to be able to invent machines. In the information era, not everyone had to make programs. Inventing was only important for specific professionals and related workers, not for the normal people. Likewise, when the AI society comes, creating or designing AI won't be asked for everyone. However, as the impact of the technology abuse was bigger in the digital era than the industrial times, the misusage of the technology will cause even worse side effects in the AI society. Therefore, we need to provide proper AI-ethics education to the students from the early days of elementary school. Then, we will have to decide what ethics to teach for the right direction. For AI-ethics education to handle AI rightly, first we need to clear the definition of the AI-ethics. For example, the following chart is the MIT Media Lab's AI-ethics learning goals among AI-ethics education in the United States.

Table 2. AI Ethics Learning Goals of MIT Media Lab (Soonyong Byun, 2020, 159)

1. Understand the basic operating principles of artificial intelligence systems.	a. Recognize various algorithms in the world, and can give examples of algorithms and computer algorithms in everyday life. b. Able to tell the three stages (input, processing, output) of the algorithm. c. Knowing that AI is a specific type of algorithm, it can be seen that it consists of a dataset, a learning algorithm, and a prediction. d. It is possible to understand AI systems in everyday life and infer what AI systems predict and potential datasets used by AI systems.
2. Recognize that all technological systems are socio-technical systems. It is understandable that socio-technological systems are not value-neutral and have political issues.	a. Understand the term optimization and recognize that the goals of man-made socio-technical systems are human-determinable. b. Be aware of social-technical systems in everyday life, and be able to distinguish between extrinsic and practical goals of social-technological systems. c. You can see the term algorithm bias that can occur in the classification process.
3. It can be recognized that there are many stakeholders in a socio-technological system and that the system can affect stakeholders in different ways.	a. Stakeholders related to the socio-technical system can be identified. b. Can reasonably explain why each stakeholder is interested in the outcome of a socio-technological system. c. It is possible to identify the value of each stakeholder in the socio-technical system. d. An ethics matrix can be built around a socio-technical system.
4. Able to apply technical understanding of artificial intelligence and knowledge of stakeholders to determine the correct goals of socio-technological systems.	a. Ethical matrices can be analyzed and the results of the analysis can be used for new goals of socio-technological systems. b. You can identify the datasets you need to train your AI system to achieve your goals. c. Functions can be designed to reflect the identified goals of a socio-technological system or to reflect stakeholder values.
5. Consider the impact of technology on the world.	a. It is possible to understand the secondary and tertiary influences caused by technology, and to know that technology is made for various stakeholders.

The MIT Media Lab focused on the ethics applied to the society and systems through the AI principles and technologies. AI cannot be ethically neutral, it can have multiple issues. Each socio-technical system is intertwined with many interests, so it can trigger the ethical or value issues. The countries and the institutions in the world are announcing various measures for AI-ethics standards or AI-ethics education. However, it hasn't been clearly set which of these should be the centered of AI-ethics education.

General AI-ethics can be divided into the ethics of AI, to treat AI, and about AI. The ethics of AI is the ethics that AI should have. That is, AI itself becomes an artificial moral actor based on semi-morality, which is the design and the application of algorithms that simulate human moral behavior(Kim Kook-hyun, 2020, 78). It relates to a question – Can AI take the responsibility or obligations when the accidents happened due to the AI's decision or malfunction. In other words, it touches the ethical field which includes the contention whether AI can become an ethical actor. It matters to the experts who develop AI, the scholars who present ethical standards, and the countries that legislate and announce the ethical standards.

This ethics needs to be formed by experts and scholars, so it will not be the central contents of the AI ethics in the school curriculums.

Next is the ethics to treat AI. The devices with AI are showing actions and reactions in a similar form to humans by learning through the technologies; machine learning, deep learning, and artificial neural networks. And the techniques, including AI's words, actions, expression methods, and avatar shape design, make AI like a human or a living creature. And because of it, humans feel more psychological bond to AI(Eunbyul Yang, Yeonju Tak, Jihun Ryu, 2021 , 2028). If AI have people feel intimate to them, they can call empathy just as animals or something people care about. In this case, AI becomes a bit different from the existing technologies or tools, and it calls some questions; Can AI be treated same as general technologies or tools? How should human act towards to AI? The answer of those questions could be the ethics to treat AI. This ethics can directly affect people’s lives more than the ethic of AI(discussed in previous paragraph), and is an ethical rule that everyone should follow. For example, when people see the video of the robot dog who is kicked to check its balancing function, if there is a demand to stop harassing or abusing of the robot dog for its right, that is the part of this ethics. There is no doubt that this ethics should be in the school curriculums, but if the technology is not improved enough to make AI be considered as a new species or an individual person, it won’t make a huge difference from the existing ethical rules. Even in the past, there is the ethics that tells us not to abuse machines or tools, especially if those resemble humans or animals. Furthermore, the ethics has been existed to treat animals because of their right. For the ethics to treat AI, the most important thing is the point of the view to see AI – Is AI be able to considered as intellectual being having dignity, so it should be discussed in the school classes. However, it can’t be the main directions of the school AI-ethics curriculums.

Finally, there is the ethics about AI. It is an ethic of how should we use AI and what we must keep in mind when using AI. Also, it considers the consequences of using AI. In the future, AI will be more used in every aspect of our lives– from controlling the simple machines to the parts closely related to life, such as autonomous vehicles and smart home appliances. Through the technology improvement, people will be able to do unimaginable things with AI. It is too early to say the era of AI arrived, but we must prepare for coming up AI society. Because of the AI’s peculiarity; widely used, convenient to invent something new, and simplify the existing progress to do something, it can cause even greater problems socially or ethically than the digital society times. Therefore, the ethics for handling AI uprightly has to be essential in school education. It will be needed for everyone regardless of class, age, occupation, etc. It must contains from the small things like writing commenting in the Internet, or SNS usage ethics, towards to the a bit more serious factors like the education for preventing the AI crimes. This ethics is also related what kind of AI or technology we should invent and develop. Thus, it is also relevant to professionals, business executives, and social leaders. That is, this ethics is necessary for all the general public who live in a AI-prevalent society, including the office workers who use AI, the experts who build AI, the businessmen who sell AI products, and the social leaders who make decisions on AI-related policies. The required level of AI-ethics could be different for each people, but the ethics about AI is an ethic to help people using AI uprightly, to prevent abusing AI to harm others, and to let the world move forward through AI.

AI-ethics for the school curriculums needs to meet several necessary and sufficient conditions. First, it should be relevant to everyone, not just some experts or limited individuals. The ethics of AI and the ethics to treat AI cannot be the center of AI ethics education, because those are more leaning towards to AI experts, academics, and social leaders. However, the ethics about AI affects the people in all classes, and could be a foundation of all ethics– the civic ethics, the professional ethics, the business ethics, and the social ethics. Therefore, the ethics about AI should form a main content of the curriculums.

Second, it should be able to expand in many directions without being limited to a certain field. As AI develops, the ethics of AI will be a very important in the future. If the singularity of AI comes as suggested by Ray Kurzweil (Kurzweil, 2007), it will be treated more crucially. Even if the singularity does not come, it will become more important the more AI does. Although, it should be defined by the AI experts, researchers, businessmen, and social leaders through the social discussion. It is not scalable to all people, so it is more suitable ethics for higher education institutions. The ethics to treat AI will also be crucial to the society after the singularity, in the sense of establishing the relationship between AI and human beings. However, until the singularity coming, it's not so different from the human-tools relations, so it doesn't have enough extensity. On the contrary, the ethics about AI as mentioned above, is more suitable for the school curriculums because of the trait to cover broader areas; from the simple using methods to even designing and improving AI.

Lastly, it need to be useful to everyone. The ethics of AI or the ethics to treat AI are the important ethics to human, technically and socially. But it's not for use to people's daily life situations. However, the ethics about AI is linked to make decisions in everyday life. In conclusion, the main contents of the AI ethic education should be the ethics applicable and useful to the public, inclusive for new and broad fields, i.e. the ethics about AI.

5. Conclusion

There is no doubt that AI will be more prevalently used in the future. Already, lots of electronic devices with AI, such as AlphaGo and Watson, are becoming more and more common in our life. These rapid development of AI causes the social, technological, and occupational changes, and it has brought confusion and renovation of the education. The students being taught in the present curriculum will live in a new society which is completely different from the one we have been through so far, will experience new changes. When they become adults, the society may become a fantastic society alike SF movies or novels. The human lives and the ethics will be changed even more than it had. Therefore, the preemptive education of the AI-ethics is demanded in schools.

Just like the social problems appeared, such as anomie phenomenon, because of the ethic education couldn't keep up with the speed of the digital era technology-development, if the AI-ethics education waits the transition towards AI society, human will not only follow the bad methods and problems before, but also welcome the new problems that we can't even imagine. Hence, we need to realize the necessity of the proper AI-ethics education and set the center of the curriculum contents. There are discussions about AI-ethics definitions in various directions, but the AI-ethics for the school curriculums should satisfy several conditions. It need to be relevant for everyone, useful for daily life decisions regardless of people's occupation or status, and expandable to the new field as the social situations changes. Therefore, the most inclusive AI ethics, namely the ethics about AI, need to be a core of the school AI ethics curriculum.

The present time cannot be said to be the era of AI. However, considering the impact that AI will bring to us and the society, the AI-ethics to use AI uprightly must be educated from the schools. Human will not be able to block the changes in the social structure caused by AI's technological advances and innovations, and also will not be able to prevent side effects completely. Thus, the importance of the AI-ethics education can't be overemphasized. The essential contents and right directions of the AI-ethics education should be set in the school curriculums, and the AI-ethics education should be started from now that the AI era hasn't come in a full scale.

References

1. Alec Ross, The Industries of the Future, social commentary(2016).
2. Byun, Sun-Yong. A Study on the Artificial Intelligence Citizenship Education. KOREAN ELEMENTARY MORAL EDUCATION SOCIETY,(67),427-445(2020).
3. Ministry of Education. Education Policy Direction and Core Tasks in the Age of Artificial Intelligence(2020).
4.Byun Sun-Yong. A Study on the Necessity of AI Ethics Education. KOREAN ELEMENTARY MORAL EDUCATION SOCIETY,31(3),153-164(2020).
5.Eunbyul Yang, Yeonju Tak, Jeeheon Ryu. The Effects of Gender Identification and Reality of Avatar on Intimacy and Intention to Use the AI Speaker. Journal of Digital Contents Society,22(12),2027-2037(2021).
6.Kim Ha Min. Revising National Curriculum of Ethics in primary education to teach AI ethics. KOREAN ELEMENTARY MORAL EDUCATION SOCIETY,(73),1-27(2021).
7.Kim Kukhyeon. Ethics of Artificial Intelligence and Moral Education. Brain, Digital, & Learning,10(1),77-92(2020).
8.Kim Tae-Chang, Byun Sun-Yong. A Study on the Necessity and Content Composition of AI Ethics Education. Journal of AI Humanitie, 8, 71-104(2021).
9.Korea Foundation for Science and Creativity.AI Education Platform Planning Policy Research Results Report. Korea Foundation for Science and Creativity(2020a).
10.Korea Foundation for Science and Creativity. Elementary and Secondary AI Education Basic Survey Report. Korea Foundation for Science and Creativity(2020b).
11.Park Hyoung Bin. Neuroscience of Human Intelligence and AI Ethics in Moral Education Curriculum Development. KOREAN ELEMENTARY MORAL EDUCATION SOCIETY,(70), 277-318(2020).
12.Ray K. The singularity is near: when humans transcend biology (Jang, S. H & Kim, M. N Trans.). Seoul: Kimyoungsa.(2007)
13.World Economic Forum. "The future of jobs report 2018." Geneva: World Economic Forum, 2018.

ICRES 2022: 7th International Conference on Robot Ethics
and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.012

BIAS AND FAIRNESS IN AI

WONSUP JUNG

School of Liberal Studies, Kyungnam University,
Changwon, Kyungnam 51767, South Korea
E-mail: wonsupj@kyungnam.ac.kr
www.kyungnam.ac.kr

This paper is focused on artificial intelligence bias. As various artificial intelligence algorithms are used in public domains such as taxation, policing, and judicial law, the debate over bias is intensifying at home and abroad despite their effectiveness. Chapter 1 introduces the main issues related to AI bias. Chapter 2 examines the reasons and types of artificial intelligence bias in light of the characteristics of the algorithm itself. In Chapter 3, I will examine some of bias issues of predictive policing algorithms that have been introduced in the United States and are producing controversy.

Concluding my presentation, I suggest several questions for our session as follows: does artificial intelligence make better judgments than humans? Are artificial intelligence algorithms the best and fairest solution to our public affairs?

1. Characteristics of Artificial Intelligence Bias

On October 6, 2020, South Korea's antitrust regulator accused Naver Corp., the nation's biggest search engine, of manipulating search algorithms in favor of the company's online shopping site and imposed a 26.7 billion-won ($22.9 million) fine. The fine, announced by the Korea Fair Trade Commission (KFTC), is the nation's first that the regulator has levied on a technology platform operator for making algorithmic changes that favor a certain business.[a] On the same day, the US House of Representatives also released a report accusing the abuse of dominance in the digital market against four major online companies, including Amazon, Apple, Facebook, and Google.[b]

Naver, which has virtually dominated the digital market in Korea, has already been a subject of constant debate over the bias of news distribution, so this situation has not little implications for the fairness of digital giants.

The solution direction seems clear to some extent in that it is a basic rule violation that the artificial adjustment of the company insiders caused inequality in the market. In this case, as it is a typical deviation or tyranny of large corporations, it is a path that can be solved in many traditional ways, such as punishing offenders, emphasizing professional ethics, or restructuring various laws and systems that ensure fair competition in the market while protecting whistle blowers.

In March 2016, we were shocked by Microsoft's chatbot Tay in a completely different way. Tay, which is known to have learned with big data, has poured out all sorts of hate speeches about the socially underprivileged, such as women, black people, and Jews, within a day after launch. The shock from Tay was even greater because of the tremendous power that Google Deep Mind's AlphaGo showed in the game against Lee Se-dol.

[a] The Korea Herald(2020, Oct. 06), "Naver faces 26.7b-won fine, accused of manipulating algorithms." http://www.koreaherald.com/view.php?ud=20201006000715&ACE_SEARCH=1.

[b] Jerrold Nadler, et al. Investigation of Competition in Digital Markets (Majority Staff Report and Recommendations, Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary).

However, two issues emerged as it became known that Tay's remarks were the result of some Twitter users actively learning Tay in a specific direction.

The first issue is the possibility that artificial intelligence technology will be systematically abused after being hacked by a specific group. The discussion on ethics and governance regarding Artificial Intelligence is exploding, and various guidelines have been released by EU Council, US government, Asian Countries, various NGOs, and Big Hi-tech Companies.[c]

The second issue is that existing social prejudices or stereotypes may be represented and/or justified or deepened through machine learning technology and artificial intelligence. Generally, whenever a new technology was introduced, we have sincere hopes that the technology would solve existing troublesome problems, as well as concerns that if the technology was extended to other areas, it would derive new problems. Alphgo is fundamentally based on mathematics. In contrast, our language proceeds in a specific socio-cultural context called “here and now”.

The Tay case clearly showed that the “bias” that can emerge when artificial intelligence technology is expanded far exceeds the level of concern for scientists and engineers. This is because the language Tay learned is deeply related to the biases of our society itself.

Moreover, human decisions can be made different from one person to another due to conflicts of interest, cultural relativity, limitations of information, lack of judgment, differences in values, etc., or may differ from person to person on the same issue, and in some cases, there may be errors. In terms of the diversity of values, making such different judgments may be desirable in some cases, and we are well aware of these points.

However, it is not easy to think that technology, especially artificial intelligence technology, will have a conflict of interest, cultural relativity, or differences in values. Thus, even on socially controversial issues, it can be more tempting to delegate troublesome decisions to AI, expecting that decisions made by artificial intelligence technology will be fairer than those of natural persons.

As a result, a number of algorithms have already been used to suggest improved decisions based on data accumulated in various fields such as taxes, loans, policing, and entrance examinations for a long time. Last summer, the British government, which was unable to take the high school graduation test due to the corona crisis, tried to give grades using artificial intelligence technology.[d] This was a happening as it was argued that public school students, mainly the low-income class, were at a disadvantage compared to the private school students, who were children of the wealthy class, as predicted by artificial intelligence. The remarks of Helena Webb, a senior researcher at the Department of Computer Engineering at Oxford, that these results "can be argued to be fair at the national level, but have completely lost fairness on an individual basis" clearly raises another important issue regarding the use of artificial intelligence technology.[e]

In this article, I would like to examine the bias that has begun to emerge as artificial intelligence is used in social issues such as predictive policing.

2. Types of Algorithmic Bias

As is well known, after the 2000s, methods such as deep learning, which dramatically improved the performance of existing neural networks and machine learning, were introduced into the

[c] In 2018, more than 50 celebrities from around the world who are leading AI technology announced a declaration refusing to collaborate with KAIST, raising concerns that KAIST's AI research could lead to research on killer robots. http://news.kmib.co.kr/article/view.asp?arcid=0923929416.

[d] https://www.axios.com/england-exams-algorithm-grading-4f728465-a3bf-476b-9127-9df036525c22.html.

[e] http://www.hani.co.kr/arti/international/europe/959055.html.

field of artificial intelligence and combined with big data. The efficiency and accuracy of the system have increased remarkably, and the application area has expanded.

However, Google Photo' classified black faces as 'gorillas' raised questions about the accuracy of artificial intelligence algorithms, and various social debates including the possibility of racial discrimination. At the same time, the question was raised about the compass algorithm, which is called an artificial intelligence judge. The debate is still seriously ongoing as the claims that the Compass algorithm, which has been widely used in judgments such as sentence, parole, and bail in US courts and prisons, has emerged against blacks.[f]

However, the problem is that these biased results can come from the following reasons without intentional intervention or artificial adjustment.

First, biased results can emerge because of the numerous correlations inherent in big data. In other words, when my spending patterns, learning patterns, Internet search patterns, hospital visit patterns, and phone usage patterns are combined, there can be many correlations.

When machine learning algorithms that learn themselves from patterns are combined, information about me is newly constructed and predictions about my behavior become possible. As a result, decisions made by algorithms combined with big data can deliberately or unintentionally produce discriminatory results.

First of all, deliberate discrimination can occur in algorithmic decision making, even if sensitive categories such as gender, race, and rich and poor are not directly included.

Second, biased results can emerge in the four stages of data mining. Data mining consists of four steps:

First, target variable definition
Second, data labeling and gathering,
third, feature selection
and fourth, decision making based on modelling

In this way, not only when discrimination is intended, but even though there is no explicit intention of discrimination, it is sufficient that bias will be involved in each step of data mining.

If so, is it possible to improve the problem of algorithmic unfairness by minimizing the potential for bias, discrimination, and unfairness at each of these steps?

At first glance, it is a valid reasoning, but several researchers point out that attempts to ensure algorithmic fairness may raise new problems. This is because efforts to solve the problem of algorithm fairness may conflict with efforts to estimate or predict unknown information as accurately as possible, which is the original purpose of big data algorithms. For example, if more detailed data is collected to increase accuracy, the combination of detailed data can eventually lead to a specific layer.[g] The problem of algorithm fairness has a trade-off with the problem of algorithm accuracy.

3. Bias in Predictive Policing

According to a report published by RAND Corporation, there are four types of predictive policing:

predicting a crime,
predicting a criminal,

[f] Angwin, Larson, Mattu, and Kirchner, 2016.
[g] Selbst, 2018: 137; cf. Williams et al., 2018.

predicting the identity of a criminal (from past crimes),
and predicting a victim.[h]

These crime prediction policing usually consists of four steps:

First, data collection step,
second, collected data analysis and prediction step,
third, crime response strategy planning step,
and fourth, criminal response step

In addition, this process continuously “feeds back” and improves the accuracy of predictions, thereby increasing the efficiency of public safety. What is important in predicting crime is where and when the crime is expected. David Weisburd argued that the paradigm of crime should shift from a people-centric paradigm to a place-centric paradigm. A company called PredPol was founded in 2012 to commercialize this technology. The company's crime prediction algorithm, Fred Paul, has garnered great media attention since its launch.[i]

Even if the prediction is not 100% accurate, predictive algorithms have been evaluated to have several advantages. First of all, the algorithm made it possible to focus the patrol power of the police in a specific area or situation. And it provided data to analyze in which circumstances crime is more likely to occur. It allowed them to plan measures to prevent crime more effectively, and could even help police use specific strategies.[j]

However, problems with the Fred Paul algorithm soon emerged.[k] Above all, a problem could be raised about the poor quality of data used by the algorithm. Most seriously pointed out was the criticism that these programs violate the right to citizenship, particularly the privacy set out in the Fourth Amendment. Indeed, within a few years, this criticism poured out from academia, legal circles, and human rights groups.

O'Neill criticized that it was wrong to include misdemeanor crimes in the data when creating the Fred Paul algorithm.[l] As data on misdemeanor crimes were included, areas where many blacks lived were classified as misdemeanor areas from the beginning. Police patrol these areas, and then they stop and check people who appear suspicious. Among them, people with drugs or suspicious objects were arrested. These arrests are fed back into this algorithm and function as data, increasing the success rate of the Fred Paul algorithm. This in turn leads the police to intensive patrol areas where black people live.[m]

Selbst (2018) pointed out that Fredpol's prediction was problematic in four ways.

First, the problem of bias arises in the process of computerizing crime. Crimes come in a variety of ways and there are many types, but prediction programs focus on the crimes that computers can easily recognize.

Second, bias can occur in the process of 'training' big data. Algorithms such as Fred Paul used a racially biased reference value, and it is highly likely that the area where many blacks or

[h] Perry et al., 2013.

[i] The Economist (2013. 7. 20), An Internet Archive page of Kent Police (2013. 12. 3) ‘‘Predictive Policing day of action targets burglars,’’ archived on May 2nd, 2014.

[j] Schlehahn et al,, 2015.

[k] Perry et al., 2013: 12-13.

[l] Cathy O’Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,* 2016.

[m] O’neil, 2017.

Hispanics lived from the beginning were identified as criminal areas and concentrated patrol power there.[n]

Third, bias arises in the process of selecting specific features of data.

The final bias is the imbalance between 'accuracy' and 'fairness'. Predictions can be more accurate, but fairness is more compromised. Legal scholars believe that big data and predictive algorithms are leading to further weakening of the 4th Amendment to the U.S. Constitution, which was ineffective against discrimination such as racism.[o] So Selbst proposed a way to limit the use of police. He argued that police using crime prediction algorithms should prepare and submit “Algorithm Impact Statements” (AIS). Just as environmental impact assessment has changed the past perception that pollution and health violations are inevitable for industrial development, algorithmic impact assessment can also change the current perception that minor human rights violations are inevitable for the development of the data industry. His proposal to write an algorithmic impact assessment is a proposal made under the assumption that the algorithm will continue to be used. This assumption is based on another assumption that artificial intelligence algorithms will make better judgments than human judgments.

But does artificial intelligence make better judgments than humans? Are artificial intelligence algorithms better and fairer solution to our public affairs? Or is it cheaper and seemingly more efficient solution?[p]

References

1. M. Barranco and J. R. Buchler, *Phys. Rev.* **Cf22**, 1729 (1980).
2. H. Müller and B. D. Serot, *Phys. Rev.* **C52**, 2072 (1995).
3. V. Baran, M. Colonna, M. Di Toro and A. B. Larionov, *Nucl. Phys.* **A632**, 287 (1998).
4. V. Baran, M. Colonna, M. Di Toro and V. Greco, *Phys. Rev. Lett.* **86**, 4492 (2001).
1. Angwin, J. and Larwon, J., ‘‘Bias in Criminal Risk Scores is Mathematically Inevitable, Researchers Say’’, *ProPublica*, (2016). [https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mat hematically-inevitable-researchers-say]
2. Burrell, J., ‘‘How the machine ‘thinks’: Understanding opacity in machine learning algorithms.’’, *Big Data & Society*, **Vol. 3, No. 1**, (2016).
3. Ferguson, A. G., ‘‘Big Data and Predictive Reasonable Suspicion’’, *University of Pennsylvania Law Review*, **Vol. 163, No. 2**, pp. 327-410, (2015).
4. Ferguson, A. G., *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, NYU Press, (2017).
5. Greene, D., Hoffmann, A. L., and Stark, L., ‘‘Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning’’, *The 52nd Annual Hawaii International Conference on System Sciences (HICSS)*, Maui, HI, (2019).
6. Hand, D. J., ‘‘Classifier Technology and the Illusion of Progress’’, *Statistical science*, **Vol. 21, No.1**, pp. 1-14, (2006).
7. Liu, H. and Motoda., H., *Feature selection for knowledge discovery and data mining*, Springer Science & Business Media, (2012).
8. Michael, M., & Lupton, D., ‘‘Toward a manifesto for the ‘public understanding of big data’.’’, *Public Understanding of Science*, **Vol. 25, No. 1**, pp. 104-116, (2016).

[n] Selbst, 2018: 133-135.
[o] Perry et al., 2013: 13.
[p] Yoehan Oh and Sung-ook Hong (2018), “Does Artificial Intelligence Algorithm Discriminate?” *Journal of Science & Technology Srudies* 18(3).

9. Jerrold Nadler, et al., Investigation of Competition in Digital Markets (Majority Staff Report and Recommendations, Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary).
10. Noble, S. U., *Algorithms of Oppression: How search engines reinforce racism*, New York: NYU Press, (2018).
11. Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., and Hollywood, J. S., *Predictive policing: The Role of Crime Forecasting in Law Enforcement Operations, Rand Corporation*, (2013).
12. PredPol (n.a.), ‘‘Company | Management Team | About Us | PredPol Predictive Policing’’ http://www.predpol.com/about/company/.
13. Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C., ‘‘Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms’’, *In Data and Discrimination: Converting Critical Concerns into Productive: A preconference at the 64th Annual Meeting of the International Communication Association. Seattle*, WA, (2014).
14. Schlehahn, E., Aichroth, P., Mann, S., Schreiner, R., Lang, U., Shepherd, I. D. H., and Wong, B. L. W., ‘‘Benefits and Pitfalls of Predictive Policing’’, *Proceedings of 2015 European Intelligence and Security Informatics Conference (EISIC 2015)*, (2015). [https://ieeexplore. ieee.org/document/7379738].
15. Selbst, A. D., ‘‘Disparate Impact in Big Data Policing’’, *Georgia Law Review*, **Vol. 52**, pp. 109-195, (2018).
16. Toole, J. L., Eagle, N., and Plotkin, J. B., ‘‘Spatiotemporal correlations in Criminal Oense Records.’’, *ACM Transactions on Intelligent Systems and Technology*, **Vol. 2, No. 4**, pp. 1-38, (2011).
17. Weisburd, D., ‘‘Place-based policing’’, *Ideas in American policing*, **Vol. 9**, pp. 1-15, (2008).
18. Williams, B. A., Brooks, C. F., and Shmargad, Y., ‘‘How Algorithms Discriminate Based on Data They Lack’’, *Journal of Information Policy*, **Vol. 8, No. 1**, pp. 78-115, (2018).
19. Yoehan Oh and Sung-ook Hong, “Does Artificial Intelligence Algorithm Discriminate?” *Journal of Science & Technology Srudies*, **18(3)**, (2018).
20. ACM U.S. Public Policy Council and ACM Europe Policy Committee, (2017).
21. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, (2017).
22. The Economist, “Don’t even think about it - Predictive policing”, (2013. 7. 20).
23. Los Angeles Times, “Stopping Crime Before It Starts”, (2010. 8. 21); Santa Cruz Sentinel, “Modest Gains in First Six Months of Santa Cruz’s Predictive Police Program”, (2012. 2. 26); NPR, “At LAPD, Predicting Crimes Before They Happen”, (2011. 11. 26) http://www.npr.org/2011/11/26/142758000/at-lapd –predictingcrimes-before-they-happen.
24. San Francisco Weekly, “All Tomorrow’s Crime: The Future of Policing Looks a Lot Like Good Branding”, (2013. 10. 30); Santa Cruz Sentinel, “Modest Gains in First Six Months of Santa Cruz’s Predictive Police Program”, (2012.02.26).
25. An Internet Archive page of Kent Police, “Predictive Policing day of action targets burglars”, (2013. 12. 3) archived on May 2nd, 2014. https://www.axios.com/england-exams-algorithm-grading-4f728465-a3bf-476b-9127-9df036525c22.html.
26. Chammah, M. with Hansen, M., “Policing the Future: In the aftermath of Ferguson, St. Louis cops embrace crime-predicting software”, The Verge, (2016. 2. 3) https://www.thever ge.com/2016/2/3/10895804/st-louis-police-hunchlab-predictive-policing-marshall-project.
27. ‘Google Photos, ...’ https://twitter.com/jackyalcine/status/615329515909156865.
28. City of New York (n.a.) ‘‘NYC Open Data’’ https://opendata.cityofnewyork.us/ ; City of New York (n.a.) ‘‘Results matching of Policy Department (NYPD) | NYC Open Data’’

https://data.cityofnewyork.us/browse?Dataset-Information_Agency=Police+Department+%28NYPD%29.

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.016

‘GOOD’ FOR AI’S

SHIN KIM
Hankuk University of Foreign Studies, South Korea
E-mail: shinkimp@gmail.com

‘Good’ does yeoman’s job. It allows us humans to express preferences, attitudes toward a state of affairs and persons, and sometimes even objective facts of/about a state of affairs and persons. We are awfully good at figuring out, given a context, what ‘good’ means in the situation. Our abilities nonetheless do not guarantee that every individual in the same situation shall agree as to its meaning. We often talk past each other: someone might use ‘good’ to express efficiency while others might take one to express his/her liking. All these are familiar to us humans, and we do well to cope with the discrepancy. We talk to each other and balance things.

Now can robots do the same? ‘Robots ought to act ethically’ sounds okay to human ears. But how would that sound to AI? Would they be able to take the sentence and distinguish among the many different meanings that the evaluative/normative expressions? That is, would AI’s take the sentence to mean just as a human person would take its meaning? I think that it is easy to say ’no’ to these questions. “AI’s do not have internal lives of their own. No preferences, no attitudes toward a state of affairs and persons.” But I also think it important not to be hasty in answering these questions. So, opposite to the easy path, this presentation will seek the things that are required for AI to have meta-ethics of their own.

In this presentation, we shall begin by questioning if it is required for AI to understand ‘good’ appropriate to the situation that they should have internal lives of their own. I would like to suggest that it isn’t a requirement. Instead, it is one of the dogmas of anthropocentric understanding of the mind. Cleared of the misconception, it should be possible to set up and investigate the meta-ethical question for AI. This presentation investigates the very question about the meaning of the ‘good’ and the ‘bad’, the ‘right’ and the ‘wrong’ from the Robots’ perspective.

ICRES 2022: 7th International Conference on Robot Ethics
and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.017

DEVELOPMENT OF ARTIFICIAL INTELLIGENCE SERVICE ROBOT ACCORDING TO THE KOREAN ETHICAL SELF-CHECKLIST

JONG-WOOK KIM, YOUNG LIM CHOI, SANGHYUN JEONG,
Department of Electronic Engineering, Dong-A University
RS906 37, Nakdong-daero 550beon-gil, Saha-gu, Busan, Republic of Korea
Artificial Intelligence and Robotics Lab
E-mail: kjwook@dau.ac.kr, lotuswave@naver.com, j921sh@naver.com,
http://deas.donga.ac.kr

JEONG-HYE HAN
Department of Computer Education, Cheongju National University of Education
2065, Cheongnam-ro, Seowon-gu, Chungbuk, Republic of Korea
E-mail: hanjh@cje.ac.kr

This paper presents development of artificial intelligence (AI) service robots according to the Self-checklist for AI Ethics Standards announced by the Korean Ministry of Science and ICT. In this paper, we introduce the AI ethics self-checklist published in 2022 in Korea and explain development of the AI service robot to which the ethical checklist are applied. The ethical standards of the self-checklist consist of 35 items that correspond to each of the 10 core ethical requirements. In this paper, five ethical requirements applied to our home care service robot will be explained.

1. Introduction

With the development of AI after the 4th industrial revolution, AI technology has been applied to various existing services. In addition, industrial fields using AI technology were created. For example, among social media, YouTube recommends videos based on user's viewing history, and Facebook and KakaoTalk use chatbots that incorporate AI technology and are used in many business fields. Because of the popularization of AI, these technologies began to influence one's own value judgment in everyday life without realizing it. Accordingly, human beings are placed in a situation in which decision-making methods change and new values are formed.

In Korea, interest in AI started to rise in March 2016, when Sedol Lee and AlphaGo played Go [1]. Motivated by this opportunity, Korean government began to announce various policies related to AI, and in 2019, the government announced a five-year AI economic roadmap [2], suggesting that AI technology is becoming more important.

There are high concerns about the negative impact of advances in AI technology. The reason is because of gender discrimination in the Amazon AI recruitment program in 2018 [3], starting with Microsoft's racist remarks by Tay in 2016 [4] as well as the Iruda(이루다) Chabot that caused a social stir in Korea in 2020. Scatterlab’s AI chatbot Iruda has been halted due to problems with personal information leakage, remarks such as demeaning the disabled, and sexual harassment in the process of building and learning data [5]. Due to these many problems, there are many voices calling for ethical consideration from the stage of AI development. Accordingly, EU and WEF (World Economic Forum) have provided various evaluation items that can be referred to in the practice of ethical principles in the area of researching AI technology and developing AI-based products or services [6,7]. In Korea, the Ministry of Science, ICT and Future Planning of Korea developed and announced the “Self-Checklist for the Practice of AI Ethical Standards” in December 2020, through the task of “Development of AI Ethical Policy for the Realization of Human-Centered AI”.

2. AI Ethics Checklist

This AI Ethics Self-Checklist is composed of check items that can be applied universally without being specific to the major fields such as engineering, science, social science, industrial fields such as manufacturing, IT, and medical care, and the stages of planning, design and development. In addition, users are encouraged to use them according to their purpose and characteristics, and the self-checklist is not a compulsory or binding law or guideline. It respects the autonomy of companies in the relevant field and presents the checklist as a voluntary code of ethics that can respond flexibly [8].

2.1. *Purpose and Recommendation*

The purpose of the self-checklist was created to implement the 10 core ethical requirements for realizing the three principles of the national AI ethics standard in the process of developing and operating an AI system. The three major principles consist of the principle of human dignity, the principle of the common good of society, and the principle of purposefulness of technology. The principle of human dignity means “AI should be developed and utilized within the scope that does not harm human life as well as mental and physical health.” The social common good principle means that “The development and use of AI for the promotion of public interest should be able to improve the universal welfare of mankind from a social, national, and global perspective.”, and the principle of purposefulness of technology is “AI technology should be developed and utilized in accordance with the purpose and intention of being a tool necessary for human life, and the process must also be ethical.”

2.2. *Key Requirement*

As shown in Table 1, the self-checklist presents a total of 35 checklists for each of the 10 core ethical requirements: human rights guarantee, privacy protection, respect for diversity, non-infringement, public good, solidarity, data management, accountability, safety, transparency.

Table 1. Number of check lists for each ethical requirement.

Key requirements	Number of check items	Key requirements	Number of check items
Human rights guarantee	5	Solidarity	3
Privacy protection	2	Data management	3
Respect for diversity	5	Accountability	4
Non-infringement	4	Safety	3
Public good	3	Transparency	3

The use of the 10 core ethical requirements and the corresponding check items follows the voluntary guidelines reflecting the self-checklist in the process of performing tasks by members or groups who design, implement, and manufacture an AI system and maintain and manage it can be practiced in the field.

1. Human Rights Guarantee: The development and use of AI must respect the rights granted equally to all human beings and guarantee various democratic values and rights specified in international human rights law.
2. Privacy Protection: Personal privacy should be protected in the entire process of developing and utilizing AI.

3. Respect for diversity: The diversity and representation of users should be reflected in all stages of AI development and utilization, and bias and discrimination based on individual characteristics such as gender, age, disability, region, race, religion, and country should be minimized and commercialized AI should be applied fairly to all.
4. Non-infringement: AI must not be used for the purpose of inflicting direct or indirect harm to humans.
5. Public good: AI should be used not only for the pursuit of personal happiness but also for the promotion of social good and the common good of mankind.
6. Solidarity: Relations between various groups should be maintained, and AI should be used with full consideration for future generations.
7. Data management: Each data, such as personal information, must be used in accordance with its purpose, and must not be used for purposes other than the intended purpose.
8. Responsibility: Efforts should be made to minimize the damage that may occur by establishing a responsible person in the process of developing and utilizing AI.
9. Safety: Efforts should be made to prevent potential risks and ensure safety throughout the entire process of AI development and use.
10. Transparency: Efforts should be made to increase transparency and explainability to a level suitable for the situation in which AI is used by considering the trade-offs with other principles to form social trust.

3. Service Robot DORI

The care service robot DORI was designed and manufactured based on MISoRo (Moral Intelligent Social Robot) by the authors for the purpose of providing nursing care services for the elderly at home. Usually, it needs to play the role of a helper robot at home, but takes care of the elderly when there is no guardian.

Figure 1. Appearance of the care service robot DORI

Figure 1 shows appearance of DORI, which is equipped with 2D 360° LiDAR, Intel RealSense camera, and touch LCD on Turtlebot3 Waffle Pi robot platform manufactured by ROBOTIS. LiDAR is used for SLAM and navigation required for autonomous driving, and the camera is used for user identification and face recognition, posture recognition, and object recognition required for service operation. The main controller adopts Intel NUC 11 Performance kit, and the whole system operates based on ROS, and OpenCR is used to control the wheels of the robot and manage sensory data.

DORI consists of a cognitive agent architecture Soar, a chatbot engine ChatScript, a DB management system SQLite, and open source packages for voice and image processing. Figure. 2 is a flowchart showing the process of the robot agent in understanding and processing the user's utterance. The robot basically analyzes the input information according to the 5W1H principle, and understands the user's request and the surrounding situation through the result.

When the user gives a command by voice, the voice is converted to text through Google Assistant API [9] and translated into Korean via Naver Papago NMT [10] if necessary.

If the user uses profanity or behaves inappropriately, the robot warns the user that the word or action is wrong or recommends correction of it. In addition, the robot can make its own ethical decisions based on the circumstances in which the user's human rights may be violated.

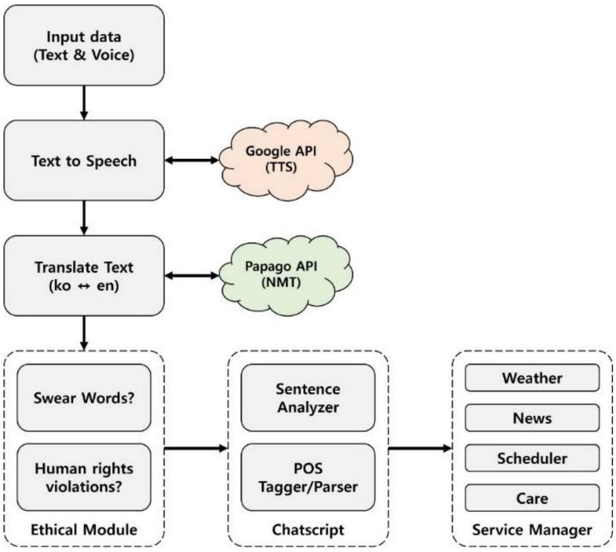


Figure 2. Flowchart of service robot DORI for speech understanding and processing

3.1. *Content and Service*

The service robot DORI basically has notification functions of weather, fine dust, news, and schedule management functions. As care service functions for elderly, DORI has service contents for enhancing cognitive activity, emotional activity, physical activity functions as well as medication guide and notification.

3.1.1. *Cognitive activity*

Cognitive function enhancement activities provide training such as guessing today's date, guessing the current year and the user's age, looking at related words and guessing names of animals and fruits, and listening to sounds and guessing transportation. The user responds to a question asked by the robot, and text cards and picture cards are displayed on the robot's screen, and the user can directly press the card to select. The robot gives the user up to three chances until they get it right, and also helps the user to find the right answer.

3.1.2. *Emotional Activity*

Emotional function promotion activities provide training such as recalling past memories by looking at pictures, listening to songs and expressing emotions. In this activity, the robot listen to the user's words, memorize them, and reflect them in the subsequent conversation. After showing a picture of the user's past, ask questions about that time. Then, after listening to music, robot asks about the favorite colors, places, and the people related with the music. At this time, when the user memorizes the user's words and has an activity to express emotions out loud, the user is trained to say what they like and express their emotions. Figure 3 is an example of conversation between user and robot during the emotional function enhancement activity.

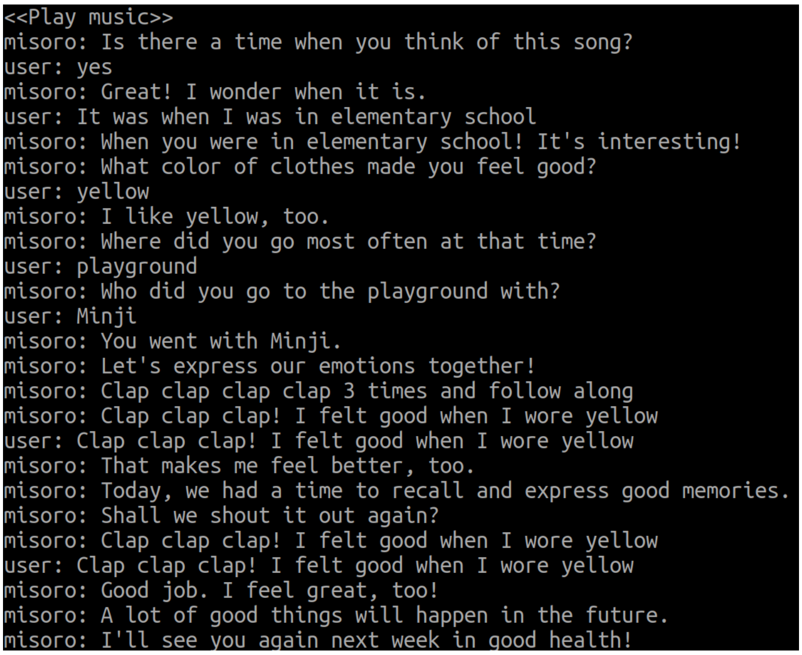


Figure 3. Example of conversation between user and robot as an emotional function enhancement activity.

3.1.3. *Physical activity*

In the activity to promote physical function, the game “Mugunghwa Flower is Blooming” is provided as shown in Figure 4. When the robot turns backwards, the activity starts, and the voice “Mugunghwa flower is blooming.” is played. When the voice playback is finished, DORI rotates in place to look at the user and then observes the user's movement. If no motion is detected, turn around and repeat this process. If motion is detected, the activity is terminated.

Figure 4. Example of physical activity program playing with ″Mugunghwa Flower is Blooming”.

At this time, posture recognition is used as a tool to observe the user's posture and movement in body function enhancement activities. The posture recognition function was implemented using the Google AI framework MediaPipe Pose [11]. MediaPipe Pose is an ML solution for high-fidelity body pose tracking that leverages BlazePose research supporting the ML Kit Pose Detection API to infer 33 3D landmarks from RGB video frames and a background segmentation mask on the whole body. Through the learned model, the user's skeleton is recognized based on the user's face from camera data.

3.1.4. *Medication guide and notification*

The medication notification service provides a function to notify the medication time at an appropriate time, and provides a function to notify the guardian through a mobile messenger whether medication has been taken or not. The medication time is stored and managed in the DB for each user, and when the set time comes, the user is identified through face recognition.

If the user's face is matched, the medicine is delivered and a medication completion message is sent to the guardian. If the user refuses to take the drug, DORI persuades the user by giving a reasonable justification. However, if it refuses more than 3 times, the service is stopped after notifying the guardian.

For face recognition, a KNN (K-Nearest Neighbor) machine learning model was used [12]. A model that has learned a user's face in advance is created, and a user is specified using real-time image processing and a learning model.

3.1.5. *Monitoring Service*

The robotic monitoring service plays a role of helping the user to live safely in the absence of his or her family or caregiver. DORI tracks and observes users in real time as soon as the service starts. It receives the image data and depth value from the camera to determine the user's location and the distance to the user. The direction of the robot is continuously controlled to align the user in the middle of the image frame, and the distance from the user is maintained at about 1~2m based on the pelvic coordinates obtained from MediaPipe Pose and depth camera. The video observed by the user is stored in the DB, and when an emergency situation occurs or the user needs help, the message is immediately notified to the guardian so that a quick response is possible. Figure 5 shows the user's location and distance between services.

Figure 5. Measurement of the user's location and distance using MediaPipe Pose and an RGBD camera.

4. Application of AI Ethics Checklist

4.1. *Cognitive and Emotional Function Promotion Activities*

Cognitive and emotional function promotion activities help users to improve their emotional stability and emotional function. Robots basically use honorifics to communicate with users in a respectful way. In addition, discriminatory remarks are not allowed according to the user's individual characteristics. This reflects the first (Human Rights Guarantee) ethical requirement, and in case of Korean, respectful words are used through English-Korean translation with Naver Papago NMT.

If the user uses profanity in the course of communication, the robot informs the user that the word is a bad one that can hurt the other person and instructs the user not to repeat it again. This is due to the fact that using profanity negatively affects not only others but also the user. Frequent use of profanity intensifies emotional anxiety and increases the likelihood of violent behavior or thinking, which in turn increases the user's aggression. This reflects the fifth (Public good) ethical requirement, and users also need to form a relationship that treats robots with respect just like humans.

4.2. *Activities to improve physical function*

In body function enhancement activities, the robot uses a camera to capture the user's movements and performs real-time posture recognition. In this case, the image data is not stored separately and is immediately removed after image processing to reflect the second ethical requirement (Privacy Protection).

4.3. *Medication Reminder Service*

In the medication notification service, the robot uses a camera to recognize the user's face for identification. The user (information subject) or guardian (third party) receives a video of the user's face to be used as learning data, and at this time, consent should be obtained from the provider according to the appropriate procedure. It also restricts data access between users by identifying users through facial recognition.

If the user ignores the medication reminder or refuses to take the medication, DORI explains the basis for medication guidance and justifiable reasons. If the user still expresses refusal, DORI says that if he or she refuses one more time, it can notify the guardian. If the user accepts taking the medicine, the medication notification service is provided as normal. However, if the user stubbornly refuses to take medicine, the current situation is communicated to the guardian and the service is stopped. This reflects the first and the fourth (Non-infringement) ethical requirements, because not performing the medication service will harm the user's health.

In addition, an accident must be prevented by providing a device or function that the user can refuse in a situation where the wrong drug is given due to an unexpected problem or dosage exceeds the daily limit. This function corresponds to the ninth ethical requirement (Safety).

4.4. *Monitoring Service*

The monitoring service tracks and observes the user in real time as soon as the service starts and informs the user to take a picture with a camera. This is a procedure for protecting the user's personal information and privacy that falls under the first (Human Rights Guarantee) and second (Privacy Protection) ethical requirements. In this case, if the user expresses a denial of service by saying "Do not follow" or "Turn off the camera", the process is immediately stopped. In consideration of the socio-economic impact corresponding to the fifth (Public good) ethical requirement, the service for elderly care robot is designed to reduce the burden on nursing and care workers and caregivers in an era of low fertility and aging.

After explaining the justification and necessity, the user's intention is checked once again. If the user says “no” again, DORI respects the user's will and stop recording the camera. It is an action based on good reason to protect the user, but should be able to abort the process if the user refuses to do so. This reflects the first ethical requirement.

Due to the characteristics of the elderly care robot, it continuously follows and observes the user, but the user may feel like being watched by the robot. To alleviate this feeling, DORI needs to maintain a distance from the user while observing for the Privacy (second) and Safety (ninth) requirements, and immediately informs the guardian of the current situation only when a dangerous situation occurs. Under normal circumstances, a third party, including a guardian, cannot know the current appearance and situation of the user.

Consequently, the authors have considered a total of five ethical requirements from the development stage of home care service robot to make a trustworthy AI robot like the following:

- 1st (Human rights guarantee): commonly crucial for robotic home care service

- 2nd (Privacy protection): removal of face and body image data
- 4th (Non-infringement): continuous request for medication
- 5th (Public good): prohibition of abusive language
- 9th (Safety): keeping distance between robot and human during monitoring

5. Conclusion

This paper presents consideration of Self-checklist for AI Ethics Standards announced by the Korean Ministry of Science and ICT in developing an elderly-care service robot. In this paper, we introduce the AI ethics self-checklist published in 2022 in Korea and explain development of the AI service robot to which the ethical checklist are applied. The authors have considered a total of five ethical requirements (first, second, fourth, fifth, and ninth ones) from the development stage of home care service robot to make a trustworthy AI robot. As future work, the authors will define detailed checklists and user experiments for these five ethical requirements which is necessary for developing and validating ethical home care service functions.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2020R1A2C1014649) and also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A4A1022059).

References

1. https://www.dongascience.com/news.php?idx=10865
2. https://www.korea.kr/news/pressReleaseView.do?newsId=156366736
3. https://www.bbc.com/korean/news-45820560
4. https://www.yna.co.kr/view/AKR20160325010151091
5. https://m.lawtimes.co.kr/Content/LawFirm-NewsLetter?serial=170076
6. The Assessment List for Trustworthy Artificial Intelligence (ALTAI), High-Level Expert Group on AI, EU, 2020
7. Companion to the Model AI Governance Framework, WEF, 2020
8. http://www.aitimes.com/news/articleView.html?idxno=141704
9. Google Assistant, https://developers.google.com/assistant
10. Papago NMT, https://developers.naver.com/docs/nmt/reference/
11. MediaPipe Pose, https://google.github.io/mediapipe/solutions/pose.html
12. Onel Harrison, Machine Learning Basics with the K-Nearest Neighbors Algorithm, https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

ICRES 2022: 7th International Conference on Robot Ethics
and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.018

THE ETHICS OF AI-POWERED VOICE CHATBOTS AND VIRTUAL REALITY FOR DIGITIAL AFTERLIVES

Eun Chang Choi
Impactive AI
28, Seolleung-ro 92-gil, Gangnam-gu
Seoul, South Korea
E-mail: eunchang.choi@aya.yale.edu

We have been facing an ethical and legal vacuum that needs to be filled in terms of using data of the deceased afterlife. Ethical aspects of the digital afterlife service are largely unexplored. The deep learning powered re-animated images and voice chatting with the deceased in virtual reality sphere are bringing psychological effect of being re-united. AI-powered virtual reality or metaverse will be a space where the bereaved could meet resurrected persons. It is highly likely that the concepts of digital grief would become a new culture. Despite of the rise of digital afterlife industry, virtual immortality raises challenging ontological questions about the demarcation between life and death.

1. The Quest for Virtual Immortality

More than three decades ago, a renowned roboticist had predicted that “More radically, we could *download* our minds directly into a body in the simulation and *upload* back into the real world when our mission is accomplished” in his book *Mind Children*.[1] Immortality, Hans Moravec suggested, is only a temporary defense against the wanton loss of knowledge and function that is the worst aspect of personal death. In order to achieve human equivalence, robots are expected to have the capacity to perform ten trillion calculations per second. The machines will have capacity to reason and ability to perceive, interact with complex environment. Most of all, the ability to reproduce biological humans into machines with minds is a sine qua non for the members of any race that aspires to immortality, he envisaged. But, his idea about ‘personal immortality by mind transplant’ has not been realized.

Moravec’s controversial quest for immortality is not anymore science fantasy or a mind-boggling glimpse in virtual world. Meanwhile, robotic engineering technology has yet reached to point in which downloads human mind into machines. But digital data the dead left can be transformed into *digital incarnation* which represents *another self* of the dead.

When it comes to uploading minds of the deceased, ‘digital remains’— the artefacts and data users consciously or unconsciously produced using email accounts, texts, social network services and other online instant messaging apps like WhatsApp, WeChat, Kakao Talk, Facebook Messenger etc.—have become increasingly available. [2] Various personal data can be used for replicating his/her social behaviour and the habit of using words in talks.

For instance, an episode of British science fiction *Black Mirro*r has a storyline that a company created a new virtual character with dead husband’s data of online communications and social media profiles. [3] The virtual character can chat with his wife in real time basis. She decided to upload more videos and photos of the husband to talk with the artificial character over the phone. ‘MyHeritage’ or ‘Deep Nostalgia’ sites allows users to produce a remarkably convincing short video with old photos of a long-lost relative or friend they uploaded. The reanimated reality can be used for family reunion as well as for satire, mockery or entertainment. A chatbot *Replika* could reproduce users identity by learning real persons’ communication style in preparation of users’ death.

Öhman & Florid suggested four types of categorization in posthumous online presence that *Digital Afterlife Industry(DAI)* offers: i) Information management service, ii) Online memorial services, iii) Posthumouse messaging service, and iv) Re-creation service. In practice, posthumouse messaging is carried out by communication tool. And re-animation is performed by artificial agents and virtual reality technology.[4]

This paper will investigate the ethical questions focusing iii) and iv). Numerous startups are already anticipating growing demand on digital mourning. In addition, onlne platforms like Facebook and Google have developed features which algorithms curate digital artefacts taken from personal online archives. What is more, Microsoft has been granted a patent that aims to revive the deceased as an artificial intelligence chatbot by preserving historical figures and living people.[5] But ethical aspects of data use and AI chatbots in digital afterlife service are largely unexplored.

2. Family Re-Union in Virtual World

Digitally resurrected persona of the dead is blurring the distinction between the quick and the dead. In the Korean documentary, entitled *Meeting You*, aired out in 2020, deep learning combined with virtual reality allowed a grieving mother to reunite with her seven-year-old-daughter, who died from blood-related diseases.[6] Digital remains of the deceased daughter included videos footages, daily conversations, family photos, preference of food, dressing style, and memory of places.

Virtual reality offered the mother with immersive experience that is different from common behaviors of mourning such as looking at photos, videos or un-animated avatars, which are incapable of interactive communications. More importantly, the mother touched her daughter with haptic gloves and head mounted display. The late daughter continuously chats with her mother on face-to-face. They were singing happy birthday song together. The mother kept weeping racked with emotion but she admitted that her daughter is dead.

Chatbots and re-animations provided the bereaved with chance to share feelings about their loved ones.Virtual experience is proved to be an effective tool of consolation and new therapeutic application on bahalf of those who had lost loved ones in grief experience.[7] With the possible long-term psychological impacts of virtual gathering on users is unclear, the mother described the re-union with re-animated daughter as satisfactory experience even though it was a very brief.

3. The Ethics of Chatting With the Dead

At a glance, chatting with the dead seems scary. But bereaved family wants to snuggle up to the lost member in virtual reality world letting alone chatting. In brief, perspectives are different. A recently developed chatbot *Replika* (https://replika.ai/) asks a series of questions to users, so that it eventually learns to mimic users' communication style.[8] Replika has been develped to give an actual voice to the digital ghosts of the deceased—seamless virtual communication even afterlife. The chatbot also has sufficient emotional responsiveness that makes users feel Replika as more of a virtual friend. At the end of day, the capability of Replika could reproduce users presence once they are dead. It deployed sequence-to-sequence methods to learns how to think and speak like a human by processing transcripts of conversations they've had during their lifetimes. Many Replika users confided in regardless of privacy. The capability of Replika makes it way more difficult to discern the voice to the digital chatbot of the deceased and real persons alive.

The first chatbot program inventor Joseph Weizenbaum raised concern over too much trust on chatbot program ELIZA.[9] It was an early natural language processing computer program created in 1960s at MIT Artificial Intelligence Lab. At the time, Weizenbaum noticed that many users formed emotional attachments to the algorithm through their interactions with Eliza. Eliza chatbot demonstrated the tendency of users who were 'willingly'or even 'complacently' falling into the illusion that conversation with the chatbot was real relationship. Weizenbaum's concern has been based on the disparity between overwhelming trust and the simplicity of Eliza design that has been simulated a Rogerian psychotherapist. Eliza became a veritable archetype of the deceitful character of AI.[10] (Natale 2021)

4. Chatbot Anthropomorphism

Chatbot anthropomorphism means the misperception that users tend to consider AI-powered voice chatbot as a real person with sentience. AI-powered voice chatting gets closer to conversational with real persons. Therefore, voice chatting with the deceased loved ones would amplify emotional empathy with a strong attachment to virtually resurrected digital persona. Then voice chatbots may be evolved beyond preserving a personality of the dead. For instance, users willingly reckon that the voice of chatbot systems could be given a legal personhood as if the deceased has never died.

Should Weizenbaum's concern be applicable to conversation via advanced chatbots with the dead? Human-extended machine cognition is considered to be a specific form of artificial intelligence. In addition, it has been also suggested that language is the basis for the brain to carry out cognition. The voice chatbots are not a system that mastered language senses or grasp the context of conversation with carefully refined words. But users unwittingly expect a higher capacity in machine judgments and communication skills (Salles et al. 2020).[11] Although voice chatbots appear to make an automated conversation, they only carry out tasks selecting one of lines in corpus datasets as designed by a human for a specific purpose with limited capabilities.

But technological settings and the purpose of chatting with the deceased are different from Eliza in 1960s. The goal of chatting with the dead in a way of digital mourning is to remember them and to get the feeling close presense of them rather than unconditional positive regard or counseling. In business interactions, chatbot anthropomorphism tends to invokes a negative effect on customer satisfaction that a limited conversation puts customers into angry emotional states.[12] But in occasion of digital mourning, there would be not much room that chatbot anthropomorphism could harm or deceive users. As far as chatting conversations are being carried out with a view to consoling and psychological effect of being re-united with the deceased family members or friends, it would be not deceitful unless conversation threads are compromised by unexpected the third party's intervention or impostering identity of the dead.

In technological perspective, seamless conversation between users and AI-powered voice chatbots resembling the deceased is the outcome of human-machine-symbiosis research. In other words, the close coupling of the deceased and voice chatbots in virtual reality indicates a move to *extended mind.*[13] It seems like a quasi human-machine symbiosis that the dead human's data cooperates with chatbot system. Unfortunatly, Weizenbaum's concern on chatbot anthropomorphism was contrasted with the vision of human-machine symbiosis.

5. Data Consent for Afterlife

When digital mourning service and virtual re-union becomes common culture, its service architects will be inevitably relying on data of the deceased. Therefore, permanent deletion of social media data or Internet service account will deprive of opportunity to build a digital memorial service or virtual immortality that bridge the gap between digital afterlife and physical world.

But not many individuals are familiar with terms of service with regard to data after death. The world-largest social media seems to have acknowledged the claim to control data afterlife. Like Facebook. Instagram's data policy offers a memorialization option but it requires proof of death such as obituary or news. Facebook. Inc., has put in place measures to control the post-mortem data on their site. The company publicly noted its afterlife policy as follows: "Memorialized accounts are a place for friends and family to gather and share memories after a person has passed away." "Content the person shared (example: photos, posts) stays on Facebook and is visible on Facebook to the audience it was shared with."[14] Otherwise, Facebook users could choose to have their account with data permanently deleted after they pass away. Verified family members also may request the removal of a loved one's account. Memorialized profiles don't appear in public spaces such as in suggestions for 'People You May Know', ads or birthday reminders.

Google's Inactive Account Manager also lets users make a plan their own digital afterlife. Users can delete Gmail messages, Blogger, Google Voice, YouTube and Google+ profile as a digital inheritance. Users can choose specific person who can control their digital assets and which of them would like to share personal data with.[15] Twitter accepts reporting the death of a user and choose between a number of options for what to do with the deceased person's account. Apple' iPhone users with iOS 15.2 operating system could indicate who can directly inherit their data after death. Apple also allows family access to iPhone data.

Unlike Facebook or Google, Naver Corp., a South Korea's Internet giant, considers user account and password as confidential data then it won't provide it to the bereaved families even if requested. It is because passwords are encrypted in a way that cannot be decrypted. Thus, it only offers backup services for publicly available data to the bereaved families due to the concern over privacy[16].

According to current laws in South Korea, personal data itself is not a property that should be inherited to immediate family. In a similar vein, the European General Data Protection Regulation no longer applies to identifiable data that relate to a personal data of the deceased.[17] Currently the control of data in social media is being handled by platform's afterlife data policy. But in case of the absence of afterlife data policy or data ethics guideline, legal disputes will arise. Prior to the dispute occurring, this ethical and legal vacuum needs to be filled regarding data ownership of the deceased. For instance, it would be ideal to set a default rule on 'prior permission ractice'. That is, anyone should set limits or conditions regarding the use of his/her own data and digital remains for afterlife, so that unauthorized third party cannot create virtual human equivalence of specific person without prior informed consent of himself/herself or ex post permission from the bereaved family.

6. Post-human Harms in Non-consensual Use

With regard to negative use of deepfakes using an AI algorithm, concerns have been raised about blackmail, intimidation, ideological manipulation, accountability and intimidation. There is growing public outcry over deepfake tools that swap real females faces – celebrities and South Korean K-Pop stars into characters of porn. AI-generated synthetic media tools and apps could strip clothes off any targeted woman photographed and create pornographic representations in a few minutes.[18] A study found that the prominence of non-consensual deepfake pornography, which accounted for 96% of the total deepfake videos online. [19]

In addition to the deepfake pornography, deepfakes are also being used in politics. In early 2022, a political party in South Korea uploaded a deepfake video of a former President Roh to Youtube platform without permission of Roh's family. The video contained AI-powered voice of Roh and it was regared as an attempt to revive the late political icon. Roh's voice portrayed a endorsement for a presidential candidate of a ruling party. But the political claim with virtual narrative got backfires from the public. The virtual endorsement by a former President Roh was regarded as illegitimate practice consisting defamation of the deceased. [20] Despite of Roh's death, his political power remains palpable in the political sphere. And it cannot be political propaganda tool or legacy property of a particular political party to which he had once been affiliated. The happening connoted technological advancement in the manipulation of political figures.

Even though deepfake technology isn't yet enough to completely deceive viewers, it presents significant ethical challenges, in particular when data of the deceased used without permission of the bereaved or beyong presumptive intention of the deceased. It is noteworthy that even posthumous events occur after death could harms the deceased**.** Some argue, however, that as long as the deceased was a public figure, there will be no room of harms in light of freedom of expression. Data of celebrities and notable figures who have passed away is not always exclusive property of their family. It is because their data—voice, photos, videos, and public speech—is already recorded online. Nonethless but, it does not necessarily mean that harm or intentional dishonoring the post-human should be allowed. For instance, after Anthony Bourdain's death, a documentary filmmaker decided to use AI to generate the voice of the late Bourdain. The artificial voice were only three particular lines in a documentary, it was enough to provoked anger of his pans.[21]

7. The Limit of Digital Immortality

Digital Afterlife Industry(DAI) will deploy State-of-the-art AI and virtual reality for digital immorality. And they will be focusing the vivid manifestation of human equivalence in digital form. Re-creation services inevitably need a range of personal data in order to generate similar conversations resembling the way in which the dead spoke it. Typically, a voice chatbot could pronounce new messages based on past data.[22] At the same time, it would raise ontological questions about the demarcation line between life and death.

Savin-Baden & Burden argues that artificial general intelligence-based virtual human could gain its own sentience and consciousness. Their viewpoints presume that coding a digital immortal would create sentience of it (Maggi Savin-Baden & David Burden 2019).[23] The possibility of autonomous presence could make us virtual sapiens. However, it is too early to predict the feasiblity of everlasting digital immortality since it hinges on the advanced Human-level AI. In the long run, deep learning technology focused on constructing virtual reality could confer the manifestation or semblance of digital immortality, but digital personas in virtual reality could not have their own sentience in virtual form.

Will should be lodged with lawyers during lifetime. Except for 5~7 percentage in South Korea, the majority of people fails to make their will before death. More than half of American adults do not leave a will. As a consequence, family members and business partners are supposed to be embroiled in complicated legal disputes. The tidal wave of resurrecting digital person afterlife would increase the expectation that the remark of virtual human equivalence could be interpreted as presumptive intention of the decesead who left no will.

References

1. Moravec. Hans, *Mind Children: The Future of Robot and Human Intelligence* (1988)
2. Lambert. Alex et al., Algorithmic memorial videos: Contextualising automated curation, *Memory Studies* Vol. 11 Issue. 2, (Nov. 2016)
3. Brooker, C. &Harris, O. *Be right back* In *Black Mirror*, Channel 4, UK (2013)
4. Öhman, C. & Floridi, L. The Political Economy of Death in the Age of Information: A Critical Approach to the Digital Afterlife Industry. *Minds Mach*. (2017)
5. U. S Patent, 'Creating a conversational chat bot of a specific person' US10853717B2.
6. VR Human Documentary, *Meeting You*, MBC http://content.mbc.co.kr/program/documentary/3479845_64342.html
7. Silvia Francesca Maria Pizzoli et al., From virtual to real healing: a critical overview of the therapeutic use of virtual reality to cope with mourning, *Curr Psychol*. (Aug. 2021)
8. Davide Sisto, *Online Afterlives: Immortality, Memory, and Grief in Digital Culture*, MIT Press (2020)
9. Joseph Weizenbaum, *Computer Power and Human Reasons* (1976)
10. Simone Natale, *Deceitful Media: Artificial Intelligence and Social Life* (2021)
11. Arleen Salles, Kathinka Evers & Michele Farisco, Anthropomorphism in AI, *AJOB Neuroscience*, 11:2, 88-95 (2020)
12. Cammy Crolic et al., Blame the Bot: Anthropomorphism and Anger in Customer–Chatbot Interactions, Journal of Marketing (2021)
13.Andy Clark, Language, Embodiment, and the Cognitive Niche, *Trends in Cognitive Sciences* 10, no. 8 (2006)
14. Facebook, What will happen to my Facebook account if I pass away? https://www.facebook.com/help/103897939701143 (Visited. April 2022)
15. Google Inactive Account Manager https://support.google.com/accounts/answer/3036546?hl=en
16. Digital Remains and Privacy Policy after Death, Naver https://bit.ly/39kwPoV
17. What is considered personal data under the EU GDPR? https://gdpr.eu/eu-gdpr-personal-data/
18. Karen Hao, A horrifying new AI app swaps women into porn videos with a click, *MIT Technology Review* (2021)
19. Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, The State of Deepfakes: Landscape, Threats, and Impact (2019)
20. Ruling party's attempt to revive late progressive icons backfires, *Korean Herald* (Feb. 7 2022)
21. Helen Rosner ,The Ethics of a Deepfake Anthony Bourdain Voice, *The New Yorker* (July 2021)
22. Carl J Öhman , Are the dead taking over Facebook? A Big Data approach to the future of death online, *Big Data & Society* (2019)
23. Maggi Savin-Baden & David Burden, Digital Immortality and Virtual Humans, *Postdigital Science and Education*. Vol. 1, 87–103 (2019)

SECTION–7

STANDARDISATION OF ROBOT SYSTEMS AND EVALUATIONS

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.022

NAVIGATION PERFORMANCE STANDARD FOR MOBILE SERVICE ROBOTS

KAI LIU and SEUNGBIN MOON
Computer Engineering, Sejong University, Seoul, South Korea
E-mail: kailiu@sju.ac.kr, sbmoon@sejong.ac.kr

Standards on the performance criteria and associated test methods play an important role in the works of international standardization. Recently, studies that intend to specify the performance criteria and test method for navigation of mobile service robots have gained many attentions. The pose characteristics is an important factor for the navigation of service robots. In this paper, we summarized the navigation performance standard, which includes the pose accuracy and repeatability of the mobile robot.

1. Introduction

The international standardization activity in the robotic field mainly consists of a few areas, such as mutual understanding, safety, performance, and interoperability [1]. The mutual understanding standards includes vocabulary and coordinate system. The safety standards deal with the safety aspect of the industrial robots and service robots and it is used as mandatory requirements in most of countries. The interoperability standards deal with the modularity of the software and hardware aspects so that these can be interchangeably used.

The performance standard is mainly used to measure the effectiveness of the robotic products in the objective and reproducible manner. While manufacturers wanted to show the superiority of their products, the customers would like to make sure the robots they are purchasing met their level of accuracy to perform the intended task of their applications. The international standard for the performance criteria and its test methods for industrial robots, ISO 9283, is historically widely utilized both by manufacturers and customers [2]. ISO 9283 mainly focused on the performance of the manipulator, such as pose accuracy and repeatability. The most widely utilized item in this standard can be the repeatability of the manipulator which effectively show the accuracy of the position and orientation of the manipulator.

As the mobile robots are widely adopted in the service robot application, the need to develop the performance standard for mobile robots is raised in the last decade. ISO 18646-1 deals with the locomotion performance of the wheeled mobile robots [3]. It includes the test methods to measure the rated speed, stopping characteristics, maximum slope angle, maximum speed on the slope, mobility over the sill, and turning width. ISO 18646-2 deals with the navigation performance of mobile service robots, such as pose accuracy and repeatability of the mobile platforms, obstacle detection, and obstacle avoidance [4]. ISO 18646-3 deals with the manipulation performance of service robots, including grasp size, grasp strength, grasp slip resistance, and use cases of opening a hinged door and opening a sliding door [5].

In this paper, we summarized the navigation performance standard of the mobile robots, ISO 18646-2. Navigation performance is usually measured by pose accuracy and repeatability, as well as the ability to detect and avoid obstacles. The criteria and related test methods are applicable only to mobile platforms that are in contact with the travel surface. To test the navigation performance three perspectives [3] are usually considered.

Firstly, the pose characteristics, which include pose accuracy and pose repeatability, is presented. Pose accuracy and pose repeatability indicate the ability of the robot to reach the command pose.

Secondly, the obstacle detection which is to determine the ability of mobile robots to detect the obstacle and measure the distance to obstacles of different geometry and material. Test parameters are based on the manufacturer specified minimum and maximum sensing ranges. The robot does not necessarily recognize the kind of obstacle.

Thirdly, the obstacle avoidance which is to determine the ability of a robot to prevent a collision with a static or dynamic obstacle, either by stopping or by conducting an appropriate evasion movement. In the case of stopping, the robot is expected to stop before physical contact between the obstacle and any part of the robot occurs. In the case of an evasion movement, a minimum distance between the obstacle and any part of the robot, as specified by manufacturer, shall be maintained.

2. Experimental conditions

All performance should be measured under normal operating conditions, and if the performance is measured in other conditions, those conditions shall be declared. Unless otherwise specifications, robot shall be tested at the rated speed and equipped with the rated load, for all tests. Additionally, the external equipment, such as landmarks, shall be supplied according to the specifications of the manufacturer and information on the external equipment, such as locations and types of landmarks, shall be provided for the navigation of mobile platforms. All test paths are parameterized with respect to the sizes of mobile platforms. Furthermore, length unit (LU) which defines as the maximum of the width *w* and the length *l* of the mobile platform, also shall be clarified in the test report. In the experiment, three types of paths are considered including straight path, rectangular path, and composite path as shown in Figures 1, 2 and 3.

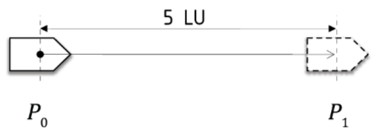


Figure 1. Straight path.

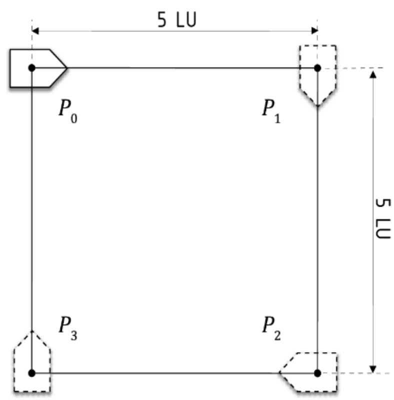


Figure 2. Rectangular path.

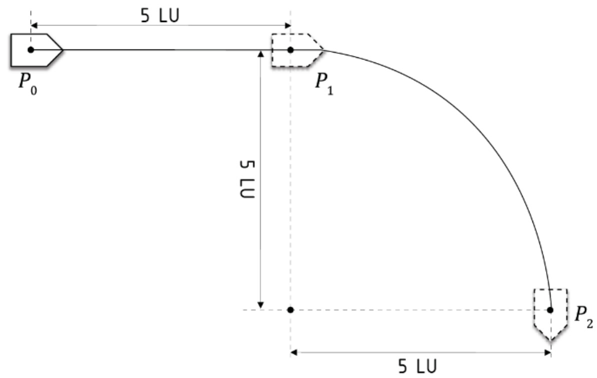


Figure 3. Composite path.

3. Pose characteristics

3.1. *Goal*

The goal of this test is to verify the performance of the navigation for service robot in terms of pose characteristics. It contains two main parts: pose accuracy and pose repeatability, which indicate the ability of the robot to reach the target pose of command.

3.2. *Pose accuracy and repeatability*

Pose accuracy is defined as the deviation between the localization poses and the achieved poses when the robot travels from the initial pose to the goal pose after n repeated visits.

Pose accuracy is divided into

1. Position localization accuracy: the difference between the position of the barycenter of the localization determined by robot and the position of the barycenter of the attained position, as shown in Figure 4.
2. Orientation localization accuracy: the difference between the average orientation of the localization determined by robot and the average of the attained orientations, as shown in Figure 5.

To Position localization accuracy (LA_p) is calculated by following formula.

$$\mathrm{LA_p} = \frac{1}{n}\sum_{j=1}^{n}\sqrt{\left(x_j - \hat{x}_j\right)^2 + \left(y_j - \hat{y}_j\right)^2} \tag{1}$$

where x_j, y_j are *x* and *y* values of the attained position of the *j-th* trial, $\hat{x}_j, \hat{y}_j$ are x and y values of localization determined by the robot of the *j-th* trial, and *n* is the number of trials.

Orientation localization accuracy (LA_o) is calculated by following formula, where the absolute value is then recast into the range (-180, +180).

$$\mathrm{LA_o} = \frac{1}{n}\sum_{j=1}^{n}\left|o_j - \hat{o}_j\right| \tag{2}$$

Here o_j is the angle of the attained pose of the *j-th* trial, $\hat{o}_j$ is the angle of localization determined by the robot of the *j-th* trial, and *n* is the number of trials.

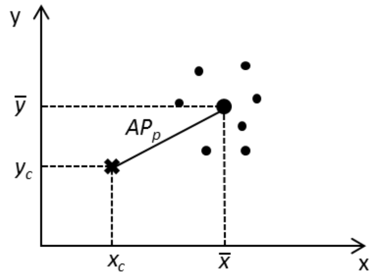


Figure 4. Position localization accuracy.

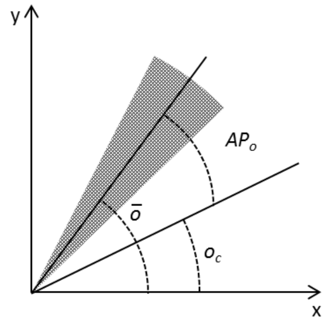


Figure 5. Orientation localization accuracy.

3.3. *Test facility*

The test facility is equipped with a measurement system which suitable for measuring position and orientation with sufficient accuracy with respect to the intended use of the robot, e.g. a 3D camera system or a laser tracker. The type and accuracy of the measurement system shall be included in the test report.

3.4. *Test procedure*

This test consists of six test configurations of a straight path, a rectangular path and a composite path (see Figure 1, 2, and 3) with no load and with the rated load. Each trial follows the procedure below.

1. The mobile platform with a specified load is placed on the initial position P0 of the respective path.
2. The mobile platform is commanded to follow the path autonomously with the rated speed.
3. When it reaches the goal position, position and orientation are measured with the external measurement system. Also, the localization determined by the robot is recorded.

The trial is repeated 30 times for each path and load condition. Position localization accuracy and orientation localization accuracy are calculated from the collected data.

References

1. Seungbin Moon, Sungsoo Rhim, Young-Jo Cho, Kwang-Ho Park, and Gurvinder S. Virk, "Summary of recent standardization activities in the field of robotics," in Robotica, February 2013, pp. 217-224.

2. Manipulating industrial robots — Performance criteria and related test methods, ISO 9283:1998.
3. Robotics — Performance criteria and related test methods for service robots — Part 1: Locomotion for wheeled robots, ISO 18646-1:2016.
4. Robotics — Performance criteria and related test methods for service robots — Part 2: Navigation, ISO 18464-2:2019.
5. Robotics — Performance criteria and related test methods for service robots — Part 3: Manipulation, ISO 18464-3:2021.

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.023

SERVICE ROBOT SAFETY STANDARDIZATION IN ISO

KWAN-WOONG GWAK
Department of Mechanical Engineering, Sejong University,
209 Neungdong-ro, Gwangjin-Gu, Seoul, South Korea
E-mail: kwgwak@sejong.ac.kr

Due to service robots' intrinsic nature of making close contact with humans, the need for securing the safety of the service robot has been increasing. As a result, the first international standard that defines the safety requirements for the service robot, ISO 13482, was published in 2014. Since then, ISO 13482 has been serving as a basis for the robot makers and certification bodies. However, service robot markets have been growing extensively and the technologies have been developed a lot, i.e. the environment around the service robot market has been changed a lot. To comply with these changes, ISO 13482 currently undergoes its revision process. In this article, brief introduction on ISO 13482 is presented and a short report on the progress of the revision is shared.

1. Introduction

Service robots interact closely with humans so its safety is critical and related risk should be reduced to an acceptable level for them to be readily accepted by the market. Therefore, there had been a need for establishing the safety requirements for the service robots agreed by the users and the manufactures. It motivated the establishment of the safety standard for the service robot, and it led to the publication of ISO 13482:2014 - Safety requirements for personal care robots[1], the first international standard establishing the safety requirements for the service robots. ISO 13482 was developed by ISO TC 299 WG 2. Since 2014, ISO 13482 has served as a guideline for the robot design, development, tests, certification by the robot manufactures, certification body and the users in the market.

Since the first publication in 2014, service robot markets have been growing extensively and segmented in various robot types. Each robot type has unique features and particular sources of hazards. However, at the time of first publication of ISO 13482 in 2014, the service robot market was at its very early stage and not many products were available on the market. Hence experiences for the real-world applications of service robots were limited. Therefore, sufficient attention was not paid to such robot type-specific requirements and robot manufacturers had difficulty applying the ISO 13482 at tis best use. Moreover, the requirements in ISO 13482:2014 did not fully reflect the limits of the reality of the robot technology and robot parts market, which are essential for the implementation of the safety requirements. As a result, ISO 13482:2014 currently undergoes its revision process. In this article, brief introduction of ISO 13482 and the revision activities in ISO TC 299 WG 2 is presented.

2. ISO 13482:2014

ISO 13482:2014 specifies the safety requirements for service robots used in personal and professional/commercial applications. The standard basically applies the philosophy of ISO

12100 specifically to the service robots, i.e., the risk assessments—hazard identification, risk estimation, risk evaluation and risk reduction—are required for the service robot use conditions. Mobile servant robots(MSR), person carrier robots(PCR), and physical assistant robots(PAR) were identified as three major robot types of service robots. Around forty hazards originate from their use were identified. To name a few, hazards related to charging battery, power failure or shutdown, robot star-up and restart of regular operation, robot shape, emission, electromagnetic interference, robot motion, mechanical instability, instability during travel, instability in case of collision, collision with safety-related objects, etc.

ISO 13482 includes the additional information on the functional safety for safety-related control systems used as protective measures. For the seven major control functions—emergency stop, protective stop, limits to workspace, safety-related speed/force control, hazardous collision avoidance, stability control, the required performance level(PLr) are suggested as in the Fig.1.

Safety functions of personal care robots	Type of robot							
	Mobile servant robot		Physical assistant robot				Person carrier robot	
	Type 1.1	Type 1.2	Type 2.1	Type 2.2	Type 2.3	Type 2.4	Type 3.1	Type 3.2
6.2.2.2 Emergency stop	d (no low risk option)		c	d	c	d	d	d
6.2.2.3 Protective stop	b	d	b	d	b	c	c	e
6.3 Limits to workspace (incl. forbidden area avoidance 6.5.3)	b^1	d	b	d	a	d	N/A	e
6.4 Safety-related speed control	b	d	b	b	b	d	c	e
6.7 Safety-related force control	b	d	b^3	e^4	a	b^5	N/A	N/A
6.5.2.1, 6.5.2.2 Hazardous collision avoidance	b	d	N/A	N/A	b	d	N/A	e^6
6.6, 6.7 Stability control (incl. overload protection)	b	d^2	N/A	c	b	d^2	b^7	d^2

Figure 1. Required performance levels for selected service robot types.

3. Revision of ISO 13482

Since its publication, ISO 13482 has served as a basis for the robot design and certification. However, there has been big changes around the service robot markets—the technical completeness of the service robot has been improved a lot due to the progress of the related technology, diverse robot types have been put into the real-world market, and competitions between robot makers have been set which requires technical completeness with low cost. To cope with these changes, ISO 13482 faces the revision.

The revision started in 2021 and the main theme of the revision is the market-relevance. ISO TC 299 WG 2 established a principle that the revision should reflect the current market only, not considering the non-existing yet to come robot types. From this perspective, the detailed classification of specific robot types is being identified and discussion on which robot types are to be included in the revised version is being made.

In terms of the structure of the revised standard, the committee are determined to separate the safety requirements that are common to various robot types and that are specific to each robot types as general safety requirements and particular safety requirements. In the particular safety requirements, additional safety requirements originate from the particular robot features are to be addressed.

Currently, the criteria by which robot type can be classified are being discussed as a major issue. In the latest e-meeting of WG 2 in April 2022, the 'function' was suggested as a key criterion.

By which, mobile robots, stationary robot and wearable robots were proposed as the current robot classification.
The revision is being processed by three subtask groups that are to establish the initial draft for three specific robot types—MSR, PCR, PAR. At the moment, initial drafts from three subtask groups are collected and it will be soon spread for the committee members for their comments. Regarding the person carrier robots(PCR), as most of the PCRs on the market are human piloted and completely lacks the autonomy, the committee temporarily determined not to consider PCR in the revised version.

Acknowledgments

This work was supported by the Technology Innovation Program (20020715) funded by the Ministry of Trade, Industry & Energy(MOTIE, Korea)

References

1. ISO 13482: 2014 - Safety requirements for personal care robots

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.024

STANDARDIZATION FOR ROBOT MODULARITY AND THE INFORMATION MODEL

HONG SEONG PARK
Department of Electrical & electronics engineering, Kangwon National University, Chencheon,24341, Korea
E-mail: hspark@kangwon.ac.kr
www.kangwon.ac.kr

Till now, many research institutes and companies have been conducting researches related to robot modularity, but there has been no talk of successful robot modularity in commercial and standard viewpoint. This is because the information provided by each institution is different according to the different perspective of modularity.

The general purpose of modularity can be listed as follows:

- Convenience of development including from design to integration
- Convenience of maintenance

Recently, the robots started to be operated based on the cloud[1], and standardization activity for the safety management system [2] was also started. It can be expected that robots operated in the cloud or safety management system will not consist of robots provided only by one robot vendors. The problem is that the following situation may occur when using robots made by various robot vendors:

- Inventory management of parts used in the various types of robots, which include mechanical and/or electrical parts and software.
- Complex management software.

: The types of parts and the software API may be different for each robot type.

: Different types of robots may have different types of communication protocols.

: Depending on the robot, the emergency management method may be different.

To solve these problems, ISO TC299 WG6 has made a standard for general requirements of robot modularity [3], where the following principles are satisfied: composability, integrability, interoperability, reusability, safety, security. In particular, since safety and security are related to each other depending on the module, risk assessment should be performed considering safety and security at the same time, and after analyzing security-related safety hazards and safety-related security vulnerability, appropriate measures must be provided. Currently, standards for robot-related information model [4-6] are being made so that the general requirement standard of ISO 22166-1 can be easily designed and implemented from a module viewpoint. The term 'information model' has been around for a long time and is being used a lot recently [7]. The definition of the term ‘information model’ is as follows [4]: abstraction and representation of the entities in a managed environment, their properties, attributes, constraints, rules, and operations, and the way that they relate to each other. Note that the information model is independent of any specific repository, usage of software aspects, communication protocol, or software platform.

In the standards for information model there have been the DMTF (Distributed Management Task Force) Common Information Model [8] and the IEC(International Electrotechnical Commission) Common Information Model (CIM)[9] used in electric power transmission and distribution. The former is an information model that is used to unify and extend the existing management standards such as SNMP and CMIP using object-oriented model in computer systems. The latter use a UML (Unified Modelling Language) model, allows application software to exchange information about an electrical network, and can be used to derive design artifacts such as XML Schema for the integration of application software.

In CD 22166-201[4], a common information model (CIM) that can be commonly applied to all modules is being developed. CD 22166-201 presents information that should be used in a common information model based on UML, which is designed based on ISO 2166-1. NP 22166-202 and - 203 are in the NP stage, and will provide an information model for software modules and an information model for hardware, respectively. These two information models are defined by inheriting the common information model.

Common information was based and modified from ISO22166-1 as shown in Table 1. In Table 1, ‘M’ and ‘O’ denotes mandatory and optional, respectively. The common information model is shown in Figure 1, which is illustrated in UML.

'Module Name' is the name representing the module. ‘Description’ provides the overview of the module, what the module is, what it does, and how it can be used. ‘Manufactures’ provides contact information for designer, developer, manufacturer of the module. ‘Examples’ provides typical use cases of the module. These 4 items are attributes of class CIM shown in Figure 1.

‘Information model version’ is the version number of information model. ‘Module ID’ is the unique identifier of the module within a robot. ‘Hardware Aspects’ and ‘Software Aspects’ are related only to composite modules. If the module is composed of two or more hardware-software modules, software modules, and/or hardware modules, their module IDs are listed in the Hardware aspects and the Software Aspects, respectively. These 4 items are attributes of class IDnType shown in Figure 1, which is one of the classes that compose of CIM.

‘Module properties’ are values that are generally used in initialization of modules. Module properties are classified into mandatory and optional ones. This information is stored into class Properties in Figure 1.

‘Inputs/Outputs’ describes names of variables for data transfer into and/or from a module, whose information is stored into class IOVariables in Figure 1. The usage of this information is similar to ROS Topics used in publish or subscribe mode.

‘Status’ describes a status of a module that is operating, whose information is stored into class Status in Figure 1.

‘Services(capabilities)’ describes information for interfaces that a module provides and utilizes for robot services, which is stored into class Services in Figure 1. The related information is the prototype of a function or method, which consists of function name, data type and name of arguments, and return type. The class Services can compose of one or more method groups, which have one or more methods.

‘Infrastructure’ lists hardware and/or software that the modules commonly uses or connects to, whose information is stored into class Infrastructure in Figure 1. Examples of the information are database type, mechanical framework, power type, and middleware type such as ROS, openRTM, and OPRoS.

‘Safety/Security’ describes the safety-related performance level and the security information provided by the module, whose information is stored into class SafeSecure in Figure 1. The safety-related information is provided as SIL(safety integrity level) IEC62061 or PL(performance level) ISO 13849-1 and the security levels 1~4 are defined in IEC 62443-4-2. If a module has several safety functions, the module provides a overall PL or SIL of the module using the combination of the performance levels of the several safety functions of the module or via the verification and validation of the module’s overall safety-related function.

Table 1. Common information for modules and the corresponding class

NO.	ITEMS	Common Information Model 2)	Information models for modules 3)			Corresponding class
			hardware-software module	hardware module	software module	
1	Module Name	M	M	M	M	CIM
2	Description	M	O	O	O	
3	Manufactures	M	M	M	M	
4	Examples	M	O	O	O	
5	Information model version	M	M	M	M	IDnType
6	Module ID	M	M	M	M	
7	Hardware Aspects	M	M	M	-	
8	Software Aspects	M	M	-	M	
9	Module properties 1)	M	M	M	M	Properties
10	Inputs	M	O	O	O	IOVariables
11	Outputs	M	O	O	O	
12	Status	M	O	O	M	-
13	Services(capabilities)	M	M	M	M	Services
14	Infrastructure	M	M	M	M	Infra
15	Safety/security	M	M	M	M	SafeSecure
16	Modelling	M	O	O	O	Modelling
17	ExecutableForm	M	O	O	M	ExecutableForm

1) It is mandatory only to those who can be influenced (set) from the outside or at least to those, which have an expected effect on other modules.
2) All kind of items for CIM are mandatory.
3) For information models such as hardware-software module, hardware module, and software module, some types of items can be omitted depending on their functionalities. In particular, for information models of hardware module and software module, the items "software aspects" and "hardware aspects" are not included, respectively.

'Modelling' provides different kinds of models for simulation and/or design purpose, whose information is stored into class Modelling in Figure 1. Paths of 3D graphic files used in simulation are stored into the class.

'ExecutableForm' provides program codes executed to achieve or support the module's purpose, whose information is stored into class ExecutableForm in Figure 1. Paths for executable files in binary and source files in python are stored into the class. Note that the former is the compiled code and the latter is the source code, which utilize the interpreter.

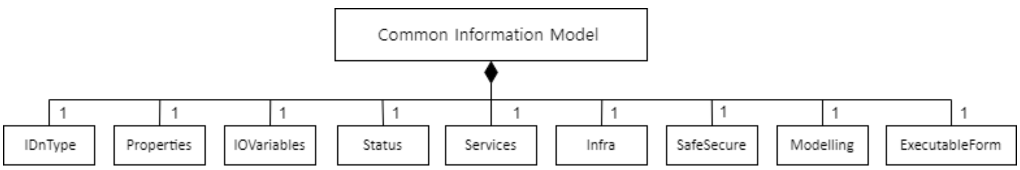


Figure 1. Class diagram of common information model

Figure 2 is the detailed class diagram for common information model, which is extended version of Figure 1 and shows the detailed attributes of classes consisting of common information model.

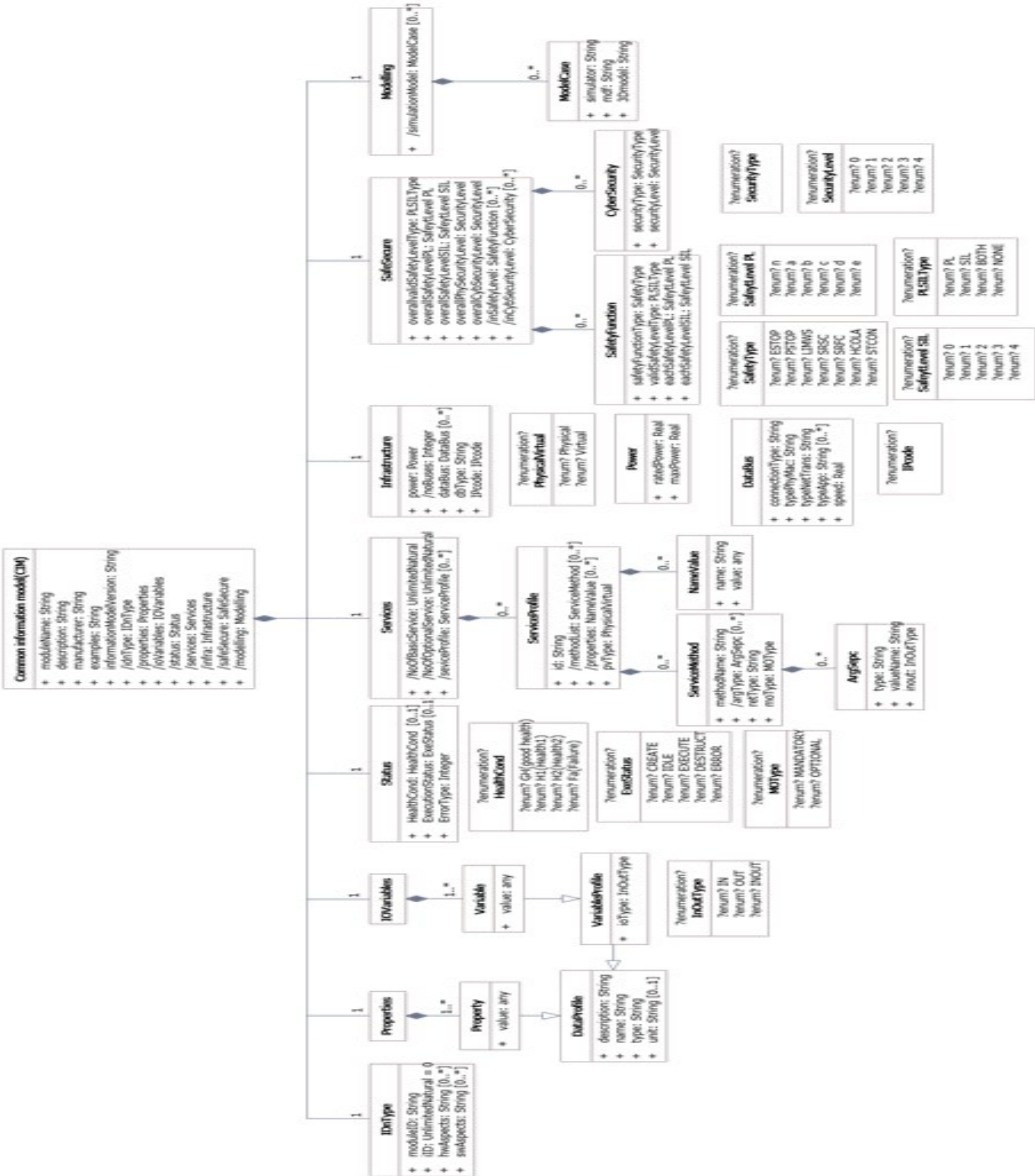


Figure 2. Detailed class diagram of common information model

This paper introduced common information for modules and the common information model, which can apply to hardware modules and software modules used in robot systems and is developing as international standards in ISO TC299 WG6.

Acknowledgments

This work was partly supported by Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (No.20006898, Standardization of Information Modeling for Service Robot) and research grant of Kangwon National University in 2022

References

1. M. Kim, S. Kim, B. Song, Y. -S. Jeong and H. S. Park, "Study on Requirements of Cloud-based Environments for Easy Development of ROS Modules," *2021 18th International Conference on Ubiquitous Robots (UR)*, 2021, pp. 48-51, doi: 10.1109/UR52253.2021.9494693.
2. ISO, ISO/TC 299/WG 7 Management system for service robots, https://www.iso.org/committee/5915511.html
3. ISO, ISO22166-1 Robotics — Modularity for service robots — Part 1: General requirements, ISO TC299/WG6.
4. ISO. ISO/CD 22166-201 Robotics — Modularity for service robots — Part 201: Common information model for modules.
5. ISO. ISO/NP 22166-202 Robotics — Modularity for service robots — Part 202: Information model for software modules.
6. ISO. ISO/NP 22166-203 Robotics — Modularity for service robots — Part 202: Information model for hardware
7. Y. Tina Lee," Information Modeling: From Design to Implementation," IEEE Transactions on Robotics and Automation, 1999
8. DMTF, https://www.dmtf.org/standards/cim
9. https://en.wikipedia.org/wiki/Common_Information_Model_(electricity)

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.025

VOCABULARY STANDARD FOR ROBOTICS IN ISO

SOON-GEUL LEE

Department of Mechanical Eng. and Integrated Education Institute for Frontier Science and Technology (BK21 Four), Kyung Hee University, 1732 Deogyeongdae-ro. Giheung-gu Yongin, Gyeonggi-do, 17104, Korea

E-mail: sglee@khu.ac.kr
www.khu.ac.kr

ISO 8373 specifies vocabulary of terms and the related definitions used in ISO documents for robotics. It defines fundamental and common vocabulary in ISO TC 299 (Robotics). The last revision on vocabulary was published in 2012, and it focused mainly to enhance the existing standard and to expand to cover the terminologies for service robots on industrial robots. ISO TC 299/WG 1 (vocabulary and characteristics) was formed to revise this standard. The final version has been developed by experts and published in 2021. Revisions are summarized in this paper.

1. Introduction

As the application area where intelligent robots are utilized expands and the global robot market is rapidly emerging in recent years, standard development organizations (SDOs), such as ISO (International Organization for Standardization), IEC (International Electrotechnical Committee), IEEE (Institute of Electrical and Electronics Engineers), OMG (Object Management Group), and ASTM (American Society for Testing and Materials), are trying to lead the standardization of the robotics field [1]. These SDOs are in a very challenging position to regulate rising domains which evolve much faster than the traditional creation cycle of a technical standard.

ISO and IEC are two major SDOs with international credibility. While ISO is developing standard for robotic area through the technical committee TC 299 (robotics) by centralized manner, IEC's robot standardization works in a distributed structure. IEC develops standards for individual robotic product by the respective TCs and SCs (subcommittees) for different application areas, such as lawn mowers, vacuum cleaning robots, or other type of household appliances. However, IEC created an Advisory Committee on Applications of Robotics Technology (ACART) to coordinate robot standardization activities within ISO and IEC.

OMG has launched the Robotics DTF (Domain Task Force) in 2005, and is currently the most active in international standardization for robot software, such as robotic middleware and its architecture. IEEE develops reports on standards by the IEEE Robotics and Automation Society (RAS). IEEE RAS has been active in standardization for more structural standards based on ontologies. ASTM has established committee E54.08(Operational Equipment) to develop standards on urban search and rescue robots. It has published standards on urban search and rescue robots, including vocabulary and performances.

The largest number of robot-related standards has so far been developed by the ISO. Originally, ISO TC184 (Industrial automation systems and integration) / SC2 which handled standards in the field of industrial robots was decided to include service robots while expanding the scope from "Robots for Industrial Environments" to "Robots and Robotic Devices" in 2006 [2]. Now, ISO TC 299 (Robotics) was newly formed by upgrading ISO TC184/SC2. It takes charge of all ISO standardization activities related to robotics, including diverse liaisons with IEC on medical robotics [3].

ISO 8373(Vocabulary) was first published in 1994 and contained mainly terms for industrial robots. The second edition was expanded to include both industrial robots and service robots and published in 2012 [4]. The last revision was published in 2021 to resolve mismatching and harmonization between different robot groups. In this paper, key vocabularies introduced in the current revision are described.

2. Basic vocabularies

The terms defined in ISO 8373 and their definitions are not only applied to other ISO standards and standards of other SDOs, but are also accepted as fundamental terms by the industry and research community in the field of robotics [5]. for example, IFR (International Federation of Robotics) uses the definitions contained in ISO 8373 "Vocabulary" when compiling statistics of industrial robots in particular countries.

As is commonly known, the vocabulary standard is essential because it must clarify the meaning of terms used in the development of a set of standards such as interfaces, performance measures and safety requirements. In the new revision, according to the emergence of new types of robots and industries, some terms and definitions that conflicts with the existing definitions were resolved by coordinating among robot groups.

The old definition of robot was described based on the mechanism point of view of an industrial robot. The definition of robot is revised to hold more broad meaning so that it can commonly be applied to service robots, but it might include some mechanisms which may not be considered as robot. To alleviate these concerns, it is specified that robots must have degree of autonomy. It was concluded that additional definitions or supplements for degree of autonomy should be defined as additional standards.

Because the revised robot definition reflects the common nature of both industrial and service robots, each group under TC 299 is able to build definitions of sub-fields based on this robot definition [6].

The term of "robotic technology" is newly introduced because it supports detail domain and reasoning of TC name "robotics". The term of "robotic device" to designate the mechanisms which might not be defined as robots according to the strict definitions of industrial robots or service robots, but utilized essential robotic technologies.

robot

programmed actuated mechanism with a degree of autonomy to perform locomotion, manipulation or positioning

NOTE 1 to entry: A robot includes the control system and interface of the control system.

NOTE 2 to entry: Examples of mechanical structure of robots are manipulator, mobile platform and wearable robots.

autonomy

ability to perform intended tasks based on current state and sensing, without human intervention

NOTE 1 to entry: For a particular application, degree of autonomy can be evaluated according to the quality of decision making and independence from human. For example, metrics for degree of autonomy exists for medical electrical equipment in IEC/TR 60601-4-1.

robotic technology

practical application knowledge commonly used in the design of robots or their control systems, especially to raise their degree of autonomy

Examples: perception, reasoning and planning algorithms

robotic device

mechanism developed with robotic technology, but not fulfilling all characteristics of robot

EXAMPLES Teleoperated remote manipulator; haptic devices, end-effector, unpowered exoskeletons, etc.

The definitions of industrial robot and service robot were revised as below, and the definitions of medical robot was newly defined based on the revised definition of robot. This means that robot is currently classified into industrial robot, service robot or medical robot, and the classification is done by its application not by its mechanical structures. This means that the robot classification can be newly classified according to the emergence of new types of robots according to the market in the future.

industrial robot

automatically controlled, reprogrammable multipurpose manipulator(s), programmable in three or more axes, which can be either fixed in place or fixed to a mobile platform for use in automation applications in an industrial environment

NOTE 1 to entry: The industrial robot includes:

- the manipulator, including robot actuators controlled by the robot controller;
- the robot control;
- the means by which to teach and/or program the robot, including any communications interface (hardware and software).

NOTE 2 to entry: Industrial robot includes any auxiliary axes that are integrated into the kinematic solution.

NOTE 3 to entry: Industrial robots include the manipulating portion(s) of mobile robots, where a mobile robot consists of a mobile platform with an integrated manipulator or robot.

service robot

robot in personal use or professional use that performs useful tasks for humans or equipment

NOTE 1 to entry: Tasks in personal use include handling/serving of items, transportation, physical support, providing guidance/information, grooming, cooking and food handling, and cleaning.

NOTE 2 to entry: Tasks in professional use include inspection, surveillance, handling of items, person transportation, providing guidance/information, cooking and food handling, and cleaning.

medical robot

robot intended to be used as medical electrical equipment or medical electrical systems

NOTE 1 to entry: A medical robot is not regarded as an industrial robot or a service robot ;.

robotics

science and practice of designing, manufacturing, and applying robots

3. Additional vocabularies

Terms such as “collaboration” and “cooperation” were revised to enhance their definition so that it can be used in precise context in other standard documents. “Collaboration” designates the situation where human and robot work together and “cooperation” is used for multiple robot cooperation. This distinction is essential in safety requirement documents, since the safety nature in “collaboration” mode is very critical.

Because “mobile platform” is more likely to be confused with other types of mobile devices, the note 3 to entry described its difference from automated guided vehicle (AGV). The note 2 to entry was added for the case that a mobile manipulator would appear as a emerging robot type in a near future.

Now the market of “wearable robot” is growing rapidly in industry and in military sides. According to these circumstances, the definition of “wearable robot” was carefully defined because it can be controversial from the point of view of “degree of autonomy”.

collaboration

operation by purposely designed robots and person working within the same space

robot cooperation

information and action exchanges between multiple robots to ensure that their motions work effectively together to accomplish the task

human–robot interaction

HRI

information and action exchanges between human and robot to perform a task by means of a user interface

EXAMPLES Exchanges through vocal, visual and tactile means.

NOTE 1 to entry: Because of possible confusion, it is advisable not to use the acronym “HRI” for human–robot interface when describing user interface.

mobile platform

assembly of the components which enables locomotion

NOTE 1 to entry: A mobile platform can include a chassis which can be used to support a load.

NOTE 2 to entry: A mobile platform can provide the structure by which to affix a manipulator.

NOTE 3 to entry: Mobile platform following a predetermined path (5.5.4) indicated by markers or external guidance commands, typically used for logistic tasks in industrial automation is also referred to as Automated Guided Vehicle (AGV) or Driverless Industrial Truck. Standards for such vehicles are developed by ISO/TC110.

wearable robot

robot that is attached to and carried by the human during use and provides an assistive force for supplementation or augmentation of personal capabilities

As the performance and safety standards of service robots were discussed, there was a discussion of definitions related to spaces which are used differently from industrial robots. A typical example is the case of “maximum space”. There was an opinion to discuss and revise this at once in connection with the “restricted space” discussed in the industrial robot group, but it was decided as in the other basic definitions as follows.

maximum space

space which can be swept by the moving parts of the robot, plus the space which can be swept by the end-effector and the workpiece

NOTE1 to entry: maximum space of the robot system can include the space which can be swept by the end-effector and the workpiece

NOTE1 to entry: For mobile platforms, this volume can be regarded as the full volume that can theoretically be reached by travelling.

working space

space which can be swept by the wrist reference point

NOTE 1 to entry: The working space is smaller than the space which can be swept by all the moving parts of the manipulator.

safeguarded space

space where safeguards are active

NOTE 1 to entry: This is sometimes described as the space within the perimeter safeguarding

NOTE 2 to entry: The safeguarded space can change dynamically.

navigation

process which includes path planning, localization, mapping, and providing the direction of travel

NOTE 1 to entry: Navigation can include path (5.5.4) planning for pose-to-pose travel and complete area coverage.

Key terms used in the modularity group were newly defined. "Component" is not a physical part used in general machines and electronics, but a conceptual element commonly applied to software and hardware. Components are able to connect, interact, or exchange resources (such as energy or data) in some way, by adhering to a standardized interface. "Module" is an independent sub-unit for a body based on "component" concept. The term "modularity" is widely used in studies of technological and organizational systems. Robots are deemed "modular", for example, when they can be decomposed into a number of components that may be mixed and matched in a variety of configurations.

component

part of something that is discrete and identifiable with respect to combining with other parts to produce something larger

Note 1 to entry: Component can be either software or hardware. A component that is mainly software or hardware can be referred to as a software or a hardware component respectively.

Note 2 to entry: Component does not need to have any special properties regarding modularity.

Note 3 to entry: A module is a component, whereas a component does not need to be a module.

modularity

set of characteristics which allow systems to be separated into discrete modules and recombined

module

component or assembly of components with defined interfaces accompanied with property profiles to facilitate system design, integration, interoperability, and re-use

Note 1 to entry: A module may have both hardware and software aspects. It may consist of other components (hardware and software) or other modules (hardware and software).

Note 2 to entry: This neither requires nor prevents the use of Open Source Software to implement parts or all of the open module's functionalities.

4. Conclusion

The recent revision of ISO 8373:2021 at TC 299/WG 1(Vocabulary and characteristics) was summarized. Experts from USA, UK, Germany, France, Sweden, Japan, and Korea has been actively engaged in the WG 1 activity during last few years. As the robot industry is widened to include both industrial robot and service robot, we needed to expand the vocabulary standard to harmonize the terminology used in both areas. We wish that the use of ISO 8373 will facilitate the development of robotic products in the market.

Acknowledgments

It was supported by Ministry of Trade, Industry and Energy under Robot Industrial Core Technology Development Project program (K_G012000921401 and 20015052) supervised by the KEIT.

References

1. S. Moon, S. Rhim, Y.-J. Cho, G. S. Virk, "International standardization activities in robotic fields", in Proc. 12th Int. workshop on Climbing and Walking Robots and Support Technologies for Mobile Machines, pp. 140-146, 2009.
2. S. Moon and G. S. Virk, "Survey on ISO standards for industrial and service robots", in Proc. ICROS-SICE Int. Joint Conf., pp. 1878-1881, 2009.
3. C. Harper, S. Dogramadzi, and M. O. Tokhi, "Development in vocabulary standardization for robots and robotic devices", in Proc. 12th Int. Conf. on Climbing and Walking Robots and Support Technologies for Mobile Machines, pp. 155-162, 2009.
4. ISO Standard, ISO 8373:2012(E) Robots and robotic devices — Vocabulary.
5. S. Moon, S.-G. Lee and K.-H. Park, "Recent progress of robotic vocabulary standardization efforts in ISO," in Proc. SICE Annual Conf., pp. 266-268, Aug. 2010.
6. S. Moon, S-G. Lee, "Revision of Vocabulary Standard for Robots and Robotic Devices in ISO", Proc. 15th Int. Conf. on Climbing and Walking Robots and the Support Technologies for Mobile Machines, Baltimore, MD, USA, 23-26 July 2012

SECTION–8

ROBOT-HUMAN-LIFE

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.026

ENCOUNTERING THE COMPLEX VALUE SYSTEMS OF THE LIFE-WORLD

ENDRE ERIK KADAR
University of Portsmouth, UK
E-mail: kadar_e@yahoo.co.uk

About a century ago, Husserl coined the term "life-world" to highlight the limitations of standard scientific methods in handling the problems of life-world (Husserl, 1936/1970). A key methodological problem in the sciences dealing with living systems is how to use of objective scientific methods for investigating the "subjective perspective" of living systems. Phenomenology was proposed and developed for this purpose but that was insufficient to solve the problem because of the complex value systems living systems rely on. Different species have different value systems (preferred habitat, hiding places from danger, favorite food source, time of mating, etc.) and there is also variability in the use of a value system within species. Human societies also have their complex value systems with differences for various social groups and for each individual. Moral principles are part of the value systems of human societies. Ethical behavior is guided by these principles but there is no evidence of unified morality across different societies even though there are some similarities across culture with respect to some abstract values such as conformity, hedonism, security (Schwartz, 1994). Differences between ethical principles of various cultures are evidenced by differences in religions. These differences create tension and possible conflicts between cultures and social groups, but globalization seemed to facilitate reducing these tensions and during the past few decades international organizations made some progress in establishing common principles in human rights, international laws, etc. Nevertheless, managing and understanding the value systems humans rely on in everyday life is a difficult task. In sociology and social psychology, empirical research of value systems relied on the assumption of static values and multidimensional scaling was their popular research method from the 1980s (Bilsky & Janik, 2010). Despite fine-tuning this method for 2-3 decades, it became apparent that the use of human value system is highly contextual and the relations between different values are not fixed (Tormos et al., 2017). However, context-dependency is not sufficient to understand the complexity of value systems. Some researchers, for instance, distinguish between fixed and fluid values (Sirris, 2020), but even fundamental human values such as human life and human freedom do not seem to be fixed. Recent pandemic experiences provided numerous examples how various countries tried to minimize mortality whilst trying to find a trade-off between human freedom and restriction of freedom to save lives. Political decision-making is only one side of the same coin. The other side is the social acceptance of and compliance with the political decisions. Countries with similar cultures (e.g., Scandinavian countries or Anglo-Saxon countries such as Australia, U.S. and UK) had different policies and the majority of population accepted their government's policies even though there were different associated risks to human life (different levels of expected and actual mortality). These differences seem to suggest fluidity of the core value of human life. Nevertheless, it was shocking to realize that large number of people showed rigidity in favoring freedom at the cost of human lives (e.g., in mask wearing, compliance with lock down rules, accepting the need for vaccination, etc.). Robotic researchers should be keenly aware of these complexities of value systems whilst designing autonomous agents that fit into the life-world. In particular, it is imperative to take the value systems of the life-world seriously in safe and ethical designs of robots and their reliance of ultimate responsibility of human controllers. However, empirical findings suggest that the fluidity of using value systems (including objective social rules and norms of behavior control) is often biased "subjective interpretation" for the benefit of the specific agent. This is in contrast with the behavior control of autonomous robots that are expected to comply with safety and moral standards. In addition to assess these issues from scientific perspectives, it is important to keep an eye and learn about the views and concerns of the public on the use of robots.

References

Bilsky, W & Janik, M. (2010) Investigating value structure: Using theory-based starting configurations in Multidimensional Scaling (Research Note), International Journal of Social Psychology, 25:3, 341-349, DOI: 10.1174/021347410792675606

Husserl, E. (1936/1970). The crisis of the European Sciences.

Schwartz, S.H. (1994). Are there universal aspects in the content and structure of values? Journal of Social Issues, 50, 19-45.

Schwartz, S. H. (2010). Values: Individual and cultural: In F. J. R. van de Vijver (Eds), Fundamental questions in cross-cultural psychology (pp.463-493). Cambridge University Press.

Sirris, S. (2020). Values as Fixed and Fluid: Negotiating the Elasticity of Core Values. In: Askeland, H., Espedal, G., Jelstad Løvaas, B., Sirris, S. (eds) Understanding Values Work. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-37748-9_11

Tormos, R., Vauclar, M. & Dobewall, H. (2017). Does Contextual Change Affect Basic Human Values? A Dynamic Comparative Multilevel Analysis Across 32 European Countries. Journal of Cross-Cultural Psychology 48(4). DOI:10.1177/0022022117692675.

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.027

TRUSTING THE HUMAN: A SOCIAL SCIENCE PERSPECTIVE

ALESSANDRA FASULO
University of Portsmouth, UK

Scientists and engineers are usually working with the standard of scientific objectivity of space-time reality based on the measures of physics. This is to ensure reliable knowledge we can trust. In contrast, human everyday thinking and behaviour control in real life settings suggest that humans experience various distortions of reality (Tversky & Kahneman, 1974; Kahneman, Knetsch, & Thaler, 1991). These distortions can be subjective but they are often systematic and labelled as biases that could undermine the reliability of everyday "know how" of humans. It is a challenge for scientist to understand why human behaviour remains mostly successful despite these distortions and biases. Moreover, human behaviour is also controlled by social norms, which consist of objective ethical rules that are often biased by subjective interpretations usually driven by self-interest. To put differently, interests of individuals or groups of individuals can be in conflict with the interest of the larger group (society) embodied in ethical rules. Empirical studies provide evidence on the specific aspects of these conflicts.

Research suggest that we are bestowed a 'hungry self', in other words, there is a built-in psychological dissatisfaction that leads individuals to pursue personal power and influence over altruistic and communitarian considerations embodied in ethical rules (Batchelor, 2011). Exchanges between humans are organised overwhelmingly on the basis of mutual advantage; cooperation is a shared moral value across societies worldwide in its dimension of reciprocity and supporting one's own group (Mauss, 1924; Curry, Mullins, Whitehouse, 2019). According to Curry et al.'s study (2019), by looking at 60 societies across the globe, deferring to superiors is also a commonly found value; together with conformism (Asch 1956) and groupthink (Janis, 1972), these can lead to discarding opinions or evidence contrary to the interests of a group or its leaders. Even without considering group pressure, studies on human's capability to form meaningful relations has been limited to around 150 people (Dunbar, 1992). On the contrary, it is known that various type of psychological processes can generate a sense of distance from other's people's suffering, or indeed to problems of general interest (Trope & Liberman, 2010): the more the distance, whether is geographical or cultural, the higher the level of "construal" that is needed to feel personal engagement.

In summary, research on cultural, social and psychological processes suggest that humans cannot be trusted in observing social norms. Thus, safeguarding the interest of human society humans need a supraordinate system of rules and principles that will form a sort of moral exoskeleton that protects the society and sanctions/punishes the violation of the norms.

References

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychological monographs: General and applied, 70(9), 1-70.

Batchelor, S. (2011). Confession of a Buddhist atheist. Spiegel and Grau; New York.

Curry, O. S., Mullins, D. A., and Whitehouse. H. (2019). Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies. Current Anthropology, 2019; 60 (1): 47 DOI: 10.1086/701478

Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. Journal of Human Evolution. 22 (6): 469–493. doi:10.1016/0047-2484(92)90081-J

Janis, I. L. (1972). Victims of groupthink: a psychological study of foreign-policy decisions and fiascos. Boston: Houghton Mifflin. ISBN 0-395-14002-1.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1991). Anomalies: the endowment effect, loss aversion and status-quo bias. J. Econ. Perspect. 5, 193–206. doi: 10.1257/jep.5.1.193

Mauss, M. (1924/ 1997). “Les changes 1. Du calme!” Le populaire, January 18, 1924. pp. 566–70 in Ecrits politiques, ed, Marcel Fournier. Paris:Fayard.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. Psychological Review, 117(2), 440–463. https://doi.org/10.1037/a0018963

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty—heuristics and biases. Science, 185, 1124–1131. doi: 10.1126/science.185.4157.1124

ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.028

ETHICAL GUIDE TO DEVELOPMENT AND DEPLOYMENT OF ROBOTS

MOHAMMAD OSMAN TOKHI
London South Bank University, UK

There has been continued effort at development of guidelines and standards for the development of ethically compliant robotics technologies. The work on robot ethics, initiated by CLAWAR Association in 2012, and continued within the robotics committee of British Standards Institution (BSI) has resulted in the publication of the British Standard BS8611:2016, as the first document of its kind worldwide. The standard provides guidance to designers, developers, and users of robots through its lifecycle from design through decommissioning within an ethical framework. The focus of the standard is the impact the robotics technology may have from an ethical perspective on humans and the society, and how the risks may be reduced. Similar initiatives have been taken elsewhere. An example is the IEEE's Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, and this has resulted in the IEEE's P7000 standards, which aim to allow organizations to demonstrate that their products conform to a high level of ethics.

The BS8611 is currently under revision so to address the emerging ethical issues of the growing robotic technology designs and applications. The revised version will include the following

- *Risk assessment*: Ethical issues are categorized into societal, application, commercial/financial, and environmental in the standard for risk assessment. Relevant ethical issues, corresponding hazards, risks, possible mitigation measures and modes of verification/validation are highlighted. The risk assessment approach of ISO12100:2010 is adopted and formulated in the framework of ethical risk assessment.
- *Ethical guidelines and measures*: The range of ethical hazards identified in the standard, such as privacy and confidentiality, Respect for human dignity and human rights, Respect for cultural diversity and pluralism, legal issues, etc. have been looked from perspective of design and corresponding design parameters are provided.
- *Verification and validation*: The standard will include detailed general information covering risk analysis, design to reduce occurrence of risks, verification via testing and verification via formal methods; mitigation; suggested methods.

Standards are provided as guidance to manufacturers and developers of technology. Their adoption is not mandatory, but rather voluntary. However, it is to the advantage of the manufacturer or developer to comply with agreed standard(s) in developing their technologies. Ethical standards, on the other hand, will involve a range of other stakeholders, such the public, end users, and regulatory bodies in complying with the guidance provided. The extent to which guidance is adhered to will impact on the future shape of the society and the character of humans in the society.

References

BS8611:2016 - Guide to the ethical design and application of robots and robotic systems.

AUTHOR INDEX

ISBN: 978-1-7396142-0-1